The Impact of Varying Knowledge on Question-Answering System

Anh Nguyen Thach Ha[1,a], Trung Nguyen Quoc[1,b], Tien Nguyen Van[2,c], Hieu Pham Trung[2,d],
Vinh Truong Hoang[3,e] and Tuan Le-Viet[3,f]

[1]Department of Information Technology, FPT University, Ho Chi Minh City, Vietnam
[2]Pythera AI
[3]Faculty of Information Technology, Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam
[1,a]anhnthse173147@fpt.edu.vn, [1,b]trungng46@fpt.edu.vn, [2,c]tien.nguyen@pythera.ai,
[2,d]hieu.pham@pythera.ai, [3,e]vinh.th@ou.edu.vn, [3,f]tuan.lv@ou.edu.vn

2024 Asian Conference on Communication and Networks (ASIANComNet) | 979-8-3503-6700-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ASIANComNet63184.2024.10811070

***Abstract*—Scale up the large language models to store vast amounts of knowledge within their parameters incur higher costs and training times. Thus, in this study, we aim to examine the effects of language models enhancing external knowledge and compare the performance of extractive and abstractive generation tasks in building the question-answering system. To ensure consistency in our evaluations, we modified the MS MARCO and MASH-QA datasets by filtering irrelevant support documents and enhancing contextual relevance by mapping the input question to the closest supported documents in our database setup. Finally, we materiality assess the performance in the health domain, our experience presents a promising result not only with information retrieval but also with retrieval augmentation tasks aimed at improving performance for future work.**

***Index Terms*—Extractive generation, Abstractive generation, Knowledge-Based Question-Answering**

I. Introduction

Recent advancements in chatbots and several developments in language modeling (LM) techniques have brought remarkable performance in natural language processing tasks. However, generating long and coherent sentences suffers from repetition, truncation, and hallucination [1,2] common in a generation not only LM but also large language modeling (LLM). To reduce this risk, [3] proposed grounded-based information retrieval from external knowledge sources; [4,5] enhancing knowledge via LLM to make the quality of response with explicit query statements. Several methods successfully compare with humans in terms of answers exploiting in selecting knowledge [6,7]. However, their success lacks case in real-life practicality as answers appear in multiple contexts when querying external knowledge [8]. Therefore, the purpose of this research focuses on the use of several linguistic models ability to understand the context, extract external knowledge, and rewrite responses more smoothly with fewer hallucinations.

Long-form question answering (LFQA) [9] introduces a new task that generates detailed and explained answers to open-ended questions from the Reddit forums "Explain like I am five years old". However, Krishna et al. [1] evaluate that at least 81% of the validation questions overlap with the training/validation data, which leads to bias in training and inference. A survey of approaches challenge in medical health introduces automatic question-answering, which has been successfully applied in various domains such as search engines and chatbots [10]. The MASH-QA [11] is a publicly available large-scale benchmark dataset different from the existing machine reading comprehension with short single-span answers for question-answering. The MASH-QA answers are extracted from multiple spans within a long context document based on questions and knowledge articles from the consumer health domain. Recently, transformer encoder models such as BERT [12] and RoBERTa [13] trained from a large corpus give the best performance when adapted to specific tasks using transfer learning among them is Dense passage retrieval (DPR) [3,14]. It has been shown to outperform traditional sparse vector space models on several benchmarks by allowing the capture of semantic similarity and handling lexical variations easily integrated with existing retrieval-reader or retrieval-generate to achieve state-of-the-art performance [3,15–17]. In addition, encoder–decoder models BART [18] and T5 [17] can be used for seq2seq tasks such as machine translation, text summarization, and question-answering for promising results.

In this paper, we present a knowledge-based question-answering system using DPR [14] to retrieve the external knowledge (in this study called support documents) and the Fusion-in-Decoder [17] takes the responsibility for generating the answer. There are two observations in our proposal 1) The large language models store vast amounts of knowledge within their parameters leading to incurring higher costs and training times. This is unnecessary with language models that force models to depend on support knowledge to control the quality and reduce the hallucination. 2) Using external document retrieval not only augments intrinsic knowledge but also grounds model outputs in a knowledge source, providing interpretability (the details in Session A). Our proposed methods can be summarized as follows:

- Building a knowledge-based Question-Answering system by using a sparse transformer-based comprising both long-form and short-form answers.

[*] Equal contribution

- The fusion-in-decoder architecture is better for extractive and abstractive generation tasks by enhanced external knowledge reducing the risk of hallucinates and making the response smoothy.
- We focus on mental healthcare with the entries of transfer learning on ELI5, MS MARCO, and MASH-QA datasets. MASH-QA is a biomedicine broad domain that covers clinical, biomedicine, consumer health, and examination.

The remainder of the article is structured as follows. Section II discusses relevant works. The key idea for the healthcare system and performing the details are explained in Section III. The outcome and the work's conclusion are then reported in Sections V to VI, respectively.

II. Related Works

The key advantages of knowledge-based question-answering tasks offer many benefits use as LM [3], augmented LM and LLM [4,5]. Digging deeper, Open Domain Question Answering introduced architecture named "Retriever-Reader" analyze the various systems that follow this architecture as well as the specific techniques adopted in each of the components as a method extractive generation on SQuAD, TriviaQA, Natural Questions, and MS MARCO datasets [19–22]. Moreover, Fan et al. [9] introduced "Retriever-Generation" for the abstractive task by augmenting external knowledge and their variant approach to the efficient transformers architecture [16,23] in several datasets such as ELI5, MASH-QA, and SaaC [9,11,24]. Su et al. [25] designed an end-to-end framework for long-form question answering that combines machine reading relevant documents and extracts salient information before generating a paragraph-length answer that is faithful to the facts.

The most fundamental distinction between the SeeKer search module [26], sparse retrieval methods such as TF-IDF [27], BM25 [28], or rewriting questions made clear context QReCC [29] is focused on exact word matching or/and frequency statistics which not reflect the correct meaning of the input question when performing the task retrieval. While the DPR module [3,14] learned embeddings from a small number of questions and passages by a simple dual-encoder framework. This allows DPR to capture semantic similarity and handle lexical variations better. Some recent techniques have also attempted to adapt knowledge through editing and tuning language model variants giving a novel perspective for solving knowledge-intensive tasks by replacing document retrievers with large language model generators. Krishna et al. [1] have demonstrated the effectiveness of the Maximum Inner Product Search to retrieve Wikipedia articles relevant to a question via a transformer model with the nearest neighbor lookup. Borgeaud et al. [15] built large language model retrieval over a database of trillions of tokens, but this method has limitations in retrieval knowledge as the database is fixed, which means would not be up to date with the latest knowledge and current events.

Perhaps the closest to our work through experiments on method and several baselines is the KILT benchmark ELI5 dataset a long-form question answering a strong abstractive task [30]. We use the "Natural Language Generation" competition track (NLGen v2.1) [22] of MS MARCO in which each query has a human-generated answer and requires using the most relevant given passage to create answers "in a way in which it could be read from a smart speaker and make sense without any additional context" as extractive. Finally, the MASH-QA dataset [11] is a publicly available large-scale dataset for question-answering, with answers extracted from multiple spans within a long context document. It is based on questions and knowledge articles from the consumer health domain. Human evaluation results further validate that our proposed framework can improve generation quality in terms of relevance and factual correctness. The details setup dataset can be found in Session A.

III. Methodology

In the following section, we decompose the question-answering system into 2 modules: Retriever, which retrieves support documents using DPR [14]; Generator, which processes each question-document pair with fusion-in-decoder [17] approach and then generates an answer. This structure is similar to the retriever-reader framework that was first introduced in DrQA [31] but instead of using a reader, we train the language model with fusion-in-decode as a generator, and these two modules can be developed independently. The framework is illustrated in Fig. 1.

A. Retriever

To access support documents, we utilize DPR to manage the Retriever module. The documents and questions are encoded as dense vector representations that are calculated using two BERT models—one for encoding questions called $DPR_q(.)$, and another for encoding documents called $DPR_d(.)$. We have N knowledge passages stored in the database, denoted as $\{d_1, d_2, \ldots, d_N\}$, and represented by allow-dimensional vector embedding, $E_d \in R^{N \times D_r}$ where D_r is the hidden dimension

$$E_{di} = DPR_d(d_i) \tag{1}$$

where $i \in \{1, 2, \ldots, N\}$.

For an input question q, $DPR_q(.)$ converts the string-based question to low-dimensional vector embedding as well, E_q:

$$E_q = DPR_q(q) \tag{2}$$

We use the FAISS [32] to speed up the retrieval of support documents. The retrieval process is performed using approximate nearest neighbors. We rank and select the most relevant knowledge passages by calculating the dot-product of the two vector representations E_q and E_{di} which serve as the retrieval score.

B. Generator

When compared to the retriever-reader approach, the retriever-generator approach also consists of two stages. However, in the second stage, the retriever-generator generates free text directly to answer the question. We train the generator to produce an answer that is human like in its response, with clear grammar and expression. The generator does not simply extract start or end positions from a retrieved passage, nor does

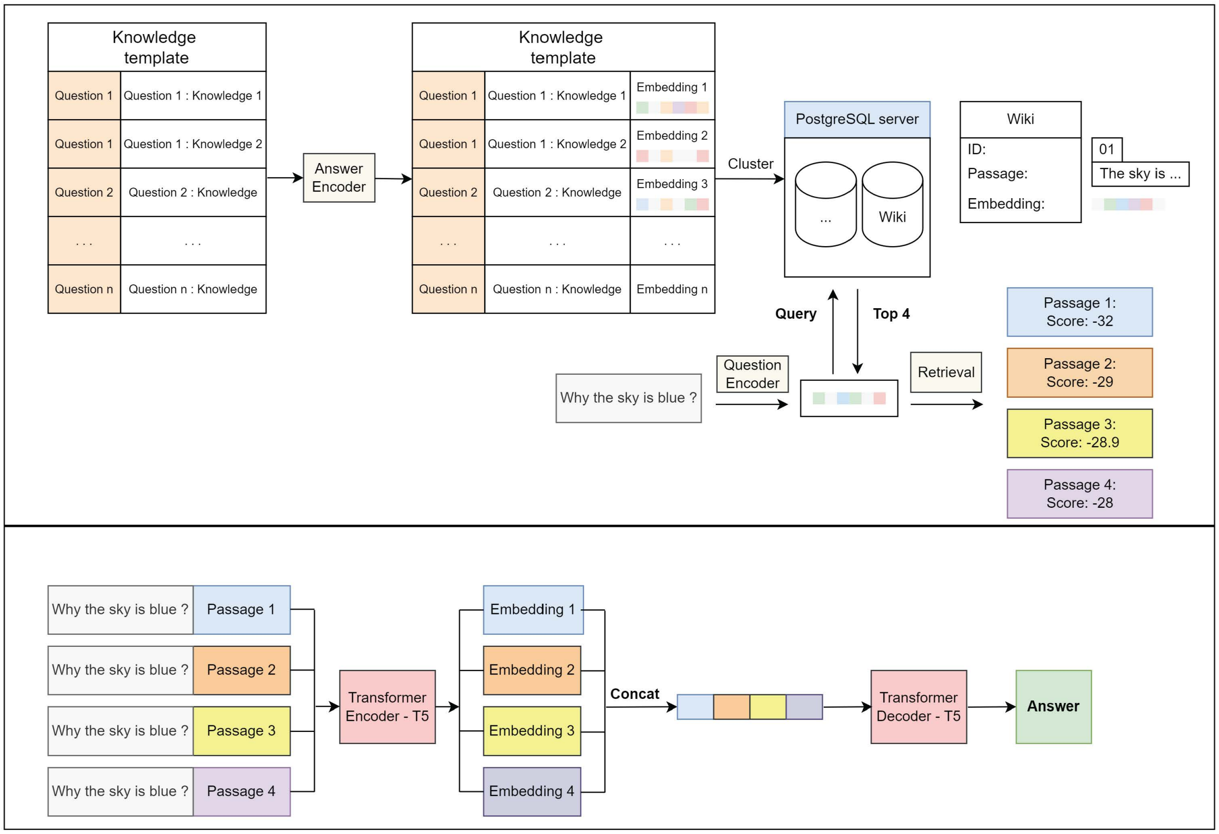

Fig. 1. FiD architecture. The figure at the top depicts the process of building a knowledge base whose data comes from a Wikipedia dump in our experiment. The knowledge base consists of text and embedding representation of each knowledge passage. Then, use FAISS to create the index for speeding up querying support documents. To query, the input question must be encoded to embedding vector, then calculated similarity score with support documents in clusters it belongs to and returns a number of knowledge passages. The figure at the bottom presents how returned support documents are paired and calculated with the question in the fusion-in-decoder approach. An input question is paired with each returned knowledge from the knowledge base. all of them are passed to the backbone model's encoder and corresponding semantic embedding vectors are generated. These embeddings are concatenated into one and passed to the decoder then generates the response.

TABLE I
THE DETAILS OF QUESTION-ANSWERING EXTRACTIVE AND ABSTRACTIVE DATASETS

Dataset	Average Length			Size			
	Question	Answer	Document	Train	Dev	Test	Total
ELI5	42.2	130.6	97.6	272,634	1,507	600	274,741
MS MARCO	6.0	13.1	93.5	453,033	2,540	1,669	457,242
MASH-QA	8.8	61.7	98.7	12,115	1,596	1,644	15,355

it generate an answer based on pieces of information already available in knowledge passages like the original FiD.

The fusion-in-decoder approach is also based on a pre-trained T5 [33]. With fusion-in-decoder, we can use more knowledge passages without them being truncated due to the token limit of any language model when concatenating all knowledge passages and a question to a single input. For each retrieved knowledge passage from retriever, a question is paired, processed independently, and later combined, then pushed to the encoder. Processing passages independently in the decoder allows us to parallelize the computation. A question and its relevant knowledge passages are separated by special prefixes such as "question: " and "context:".

Given a set of retrieved knowledge passages $\{k_1, k_2, \ldots, k_{N_\alpha}\}$, where $N_\alpha << N$. D_g is the hidden dimension of each token embedding and our T5 backbone has L encoder and decoder layers. Each string input is calculated by

$$I = f_{tokenize}(q + k_i) \tag{3}$$

$$W_i^{(0)} = f_{emb}(I) \tag{4}$$

, where $I \in R^S$

$$W_i^{(l)} = f_{t5-enc}(W_i^{(0)}) \tag{5}$$

where $W_i \in R^{S \times D_g}$, S is the max length of the input sequence. The tokenized input vector is computed embedding representation by $f_{emb}(.)$ and encoded by multiple t5 encoder layers, f_{t5-enc}. Then, the output of the last layer of the encoder is put into the decoder, $f_{t5-dec}(.)$, to compute cross-attention and generate the answer hidden state V:

$$V = f_{t5-dec}([W_1^{(l)}; W_2^{(l)}; \ldots; W_{N_\alpha}^{(l)}]) \tag{6}$$

IV. Experiments

A. Datasets

ELI5: We use the ELI5 dataset in the KILT benchmark [30], the KILT version changed the knowledge source from common Crawl to a fixed Wikipedia snapshot on August 01, 2019, that includes 5.9M articles. The total number of samples in the train, validation, and test sets is 272,634, 1,507, and 600, respectively, with the average of questions and answers being 42.2 and 130.6 words.

MS MARCO: provides a realistic setting for natural language understanding research and covers diverse topics and domains [22]. The Question Answering and Natural Language Generation task requires using the most relevant given passage to create answers "in a way in which it could be read from a smart speaker and make sense without any additional context." The method of creating this dataset involves using real Bing questions and human-generated answers. The queries are sampled from anonymized user logs and the answers are generated by human annotators based on relevant web passages. In this case, we modify the entire dataset suitable to our constraint with train, validation, and test sets are 453,033, 2,540, and 1,669 samples, respectively.

MASH-QA: The dataset is created by collecting consumer healthcare queries from a commercial search engine and matching them with relevant knowledge articles from a health website [11]. It has 34,808 question-answer pairs and 5,574 documents with answers are 67.2 words in average length, which shows this dataset serves well for long-form question-answering tasks. MASH-QA was originally used for extractive question-answering tasks. However, we approach this differently by generating answers for the questions with supporting documents from the system's external knowledge.

B. System and Parameter Settings

We used the ELI5 dataset following the instructions of the KILT benchmark kit and processed MS MARCO and MASH-QA as mentioned in Section A, which uses Wikipedia dump as the source of the documents and then employed DPR to retrieve passages for all datasets. For each question, we retrieved 30 documents and set their maximum length to 250 words. Our main training starts with the pretrained T5 model weight available in the Hugging Face Transformers library, following former research and fine-tuning the models on each dataset independently, using the Adam optimizer and learning rate of 10^{-4} (effective batch size is 64 and 128, respectively on abstractive and extractive generation tasks). We evaluate the models every 500 steps using beam search with a beam size of 4 and set a maximum answer length of 200 words.

V. Experiment Results

In this paper, we evaluate the performance of our text generation system using ROUGE-L, BLEU-1, and F1-score. ROUGE-L is a metric that assesses the similarity between two sequences based on their longest common subsequence. ROUGE-L scores range from 0 to 1 to evaluate our generated answers, with higher scores meaning better results. In addition, we also employed BLEU-1 and F1-score to assess the quality of our generated answers. The BLEU-1 score measures the percentage of individual words that perfectly match the machine output and a reference answer, with higher scores indicating greater similarity with the reference sentence. Finally, the F1-score, which is the harmonic mean of precision and recall, was used to evaluate how well the generated answer matched normalized uni-grams with the reference answer. By using multiple metrics, we aim to provide a comprehensive evaluation of our system's performance and demonstrate its effectiveness in generating high-quality text.

Our experiment involved the use of three different datasets, and we observed varying levels of success across these datasets. First of all, we compare the abstractive and extractive generation performance of our system in Tables II and III. Then, we compare the effect of the number of documents on results in Section B and abstractive and extractive generation with related and unrelated support documents in Section C. The performance of abstractive generation Table II shows that our approach ROUGE-L 29.37% and F1-score 27.78% on the dev set improve the performance with many methods while allowing customized database depends on our system using dummy Wikidataset. Besides that, in the test set, our approach keeps the original FAISS template format, and experimental results are fewer than RBG [25] on F1-score

TABLE II
Comparison of our model with several methods on a dev and test set of KILT ELI5 (the bold denotes the overall best performance)

Models	dev		test	
	ROUGE-L	**F1**	**ROUGE-L**	**F1**
T5 [30]	21.02	18.36	19.08	16.10
BART [30]	22.69	22.19	20.55	19.23
DPR+BART [30]	17.41	17.88	17.41	17.88
RAG [3]	16.11	17.24	15.50	17.10
RT+c-REAM [1]	24.40	25.60	23.20	22.90
RBG [25]	24.46	**29.04**	**24.72**	**27.52**
Ours	**29.37**	27.78	22.68	24.48

TABLE III
Performance comparison between extractive (MS MARCO, MASH-QA) and abstractive (ELI5) test sets with 30 support documents for each sample. We aim to provide a comprehensive evaluation using multiple metrics reflect assess the similarity and the quality of the response generation. Following the results, the MS MARCO dataset improves performance with the MASH-QA not only BLEU1 but also ROUGE-L and F1 scores more than 10% performance. Moreover, to balance the extractive and abstractive, we also evaluate the ELI5 dataset with ROUGE-L and F1 scores are 22.68% and 24.48%, respectively

	ROUGE-L	F1	BLEU1
MS MARCO	37.21	37.64	27.50
MASH-QA	20.69	24.60	26.63
ELI5	22.68	24.48	-

27.52% and 24.48% and improve 1.6% performance with [1] F1-score 22.9% and 24.48%. In addition, our strategic focus on the expectation gap abstractive and extractive generation, in Table III when transferring learning pretrained language model on specific tasks with mixed training the performance. Results from the MS MARCO dataset show values for 37.21% ROUGE-L, 27.50% BLEU-1, and F1-score 37.64% compared to other datasets this is likely due to several factors, including the utilization of older model architecture, and limited computing power, changes in test data, and differences in supporting documents. Similarly, the MASH-QA dataset did not yield the expected favorable outcomes, and this study marked the first attempt at utilizing this dataset for generative question-answering purposes with the results 20.69%, 24.60%, and 26.63%, respectively.

We observed more positive results with the transfer learning performance, which was characterized by several advantages on the ELI5 dataset. The supporting documents used in this dataset were of the same type as the original, resulting in reduced variability and improved model performance. Moreover, we take advantage of information retrieval with several steps of processing to enhance the analysis process's efficiency and ensure greater consistency in results in MASH-QA and MS MARCO. The proposal is able to query any number of documents as needed, facilitating data exploration and approach refinement.

VI. Conclusion and Future Work

In this study, we concentrate on evaluating the ability of the Language Model on extractive and abstractive generation tasks by enhancing external knowledge reducing the risk of hallucinates and making the response answer smoother. Moreover, we evaluate carefully the effect of relevant and irrelevant support documents on the question. Table III summarizes the entire result of our modified ELI5 and MASH-QA correspondence between model accuracy and support documents for different generation tasks. The MS MARCO with training twice the number of samples improved performance by 10% when the same hyperparameters entire training phase.

We believe that the experiment's results will be flawed in several cases. However, these knowledge gained contributions will be a premise to help the research community to exploit more in this field. The instruction tuning with meta-learning has shown strong performance on natural language generation tasks as applied in conversational AI, which trains and evaluates in dialogue context to improve multitasking which is our future work.

References

[1] K. Krishna, A. Roy, and M. Iyyer, "Hurdles to Progress in Long-form Question Answering", May 2021.

[2] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, and P. Fung, Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, page 3571730, November 2022.

[3] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, April 2021.

[4] W. Yu, D. Iter, S. Wang, Y. Xu, M. Ju, S. Sanyal, C. Zhu, M. Zeng, and M. Jiang, *Generate rather than Retrieve: Large Language Models are Strong Context Generators*, January 2023.

[5] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer, *Rethinking the Role of Demonstrations: What Makes In-Context Learning Work*?, October 2022.

[6] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, Improving Language Understanding by Generative Pre-Training.

[7] M. Li, B. Peng, J. Gao, and Z. Zhang, *OPERA: Harmonizing Task-Oriented Dialogs and Information Seeking Experience*, June 2022.

[8] S. Feng, S. S. Patel, H. Wan, and S. Joshi, MultiDoc2Dial: Modeling Dialogues Grounded in Multiple Documents. In *Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing*, pp. 6162–6176, 2021.

[9] A. Fan, Y. Jernite, E. Perez, D. Grangier, J. Weston, and M. Auli, *ELI5: Long Form Question Answering*, July 2019.

[10] Q. Jin, Z. Yuan, G. Xiong, Q. Yu, H. Ying, C. Tan, M. Chen, S. Huang, X. Liu, and S. Yu. *Biomedical Question Answering: A Survey of Approaches and Challenges*, September 2021.

[11] M. Zhu, A. Ahuja, D.-C. Juan, W. Wei, and C. K. Reddy, Question Answering with Long Multiple-Span Answers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3840–3849, Online, 2020. Association for Computational Linguistics.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *Attention Is All You Need*, December 2017.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, *Roberta: A robustly optimized bert pretraining approach*, 2019.

[14] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-T. Yih, *Dense Passage Retrieval for Open-Domain Question Answering*, September 2020.

[15] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. van den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, D. de Las Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. W. Rae, E. Elsen, and L. Sifre, *Improving language models by retrieving from trillions of tokens*, February 2022.

[16] S. Hofstätter, J. Chen, K. Raman, and H. Zamani, *FiD-Light: Efficient and Effective Retrieval-Augmented Text Generation*, September 2022.

[17] G. Izacard and E. Grave, *Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering*, February 2021.

[18] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*, October 2019.

[19] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, *SQuAD: 100,000+ Questions for Machine Comprehension of Text*, October 2016.

[20] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov, Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* vol. 7, pp. 453–466, 2019.

[21] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, *TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension*, May 2017.

[22] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, and T. Wang, *MS MARCO: A Human Generated MAchine Reading COmprehension Dataset*, October 2018.

[23] M. de Jong, Y. Zemlyanskiy, J. Ainslie, N. FitzGerald, S. Sanghai, F. Sha, and W. Cohen, *FiDO: Fusion-in-Decoder optimized for stronger performance and faster inference*, December 2022.

[24] P. Ren, Z. Chen, Z. Ren, E. Kanoulas, C. Monz, and M. De Rijke, Conversations with Search Engines: SERP-based Conversational Response Generation. *ACM Transactions on Information Systems* vol. 39, no. 4, pp. 1–29, 2021.

[25] D. Su, X. Li, J. Zhang, L. Shang, X. Jiang, Q. Liu, and P. Fung, *Read before Generate! Faithful Long Form Question Answering with Machine Reading*, March 2022.

[26] K. Shuster, M. Komeili, L. Adolphs, S. Roller, A. Szlam, and J. Weston, *Language Models that Seek for Knowledge: Modular Search & Generation for Dialogue and Prompt Completion*, March 2022.

[27] B. Das and S. Chakraborty, *An improved text sentiment classification model using tf-idf and next word negation*, 2018.

[28] S. Robertson, H. Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval* vol. 3, no. 4, pp. 333–389, 2009.

[29] R. Anantha, S. Vakulenko, Z. Tu, S. Longpre, S. Pulman, and S. Chappidi, Open-Domain Question Answering Goes Conversational via Question Rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 520–534, Online, 2021. Association for Computational Linguistics.

[30] F. Petroni, A. Piktus, A. Fan, P. Lewis, M. Yazdani, N. De Cao, J. Thorne, Y. Jernite, V. Karpukhin, J. Maillard, V. Plachouras, T. Rocktäschel, and S. Riedel, *KILT: a Benchmark for Knowledge Intensive Language Tasks*, May 2021.

[31] D. Chen, A. Fisch, J. Weston, and A. Bordes, Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1870–1879, Vancouver, Canada, 2017. Association for Computational Linguistics.

[32] J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* vol. 7, no. 3, pp. 535–547, 2019.

[33] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*, July 2020.