Retrieval-Augmented Knowledge-Intensive Dialogue

Zelin Wang, Ping Gong(✉), Yibo Zhang, Jihao Gu, and Xuanyuan Yang

School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
{wang_zelin,pgong,zhangyibo,gujihao}@bupt.edu.cn

Abstract. Large pre-trained language models have been shown to be powerful in open-domain dialogue. However, even the largest dialogue models suffer from knowledge hallucination, generating statements that are plausible but factually incorrect. Recent works, such as RAG and FiD, have introduced retrieval methods to alleviate this issue by bringing in external knowledge sources. Based on this research direction, we propose a plug-and-play method to enhance the generator's performance by introducing demonstration-based learning, which allows the generator to better understand the task. Furthermore, we propose a novel representation-interaction ranking model, RM-BERT, inspired by residual connections, which can more effectively represent the semantic information of context and document to improve retrieval accuracy. Experimental results indicate that the generator improvement method is applicable to multiple datasets and multiple models simultaneously. Additionally, we demonstrate that RM-BERT achieves performance close to BERT while significantly reducing computational overhead.

Keywords: Dialogue system · Retrieval-augmented generation · Document retrieval

1 Introduction

Large pre-trained language models, such as GPT3 [2], PaLM [4] and ChatGPT, have been shown to produce impressive performance on a range of tasks, especially in generating coherent, fluent and human like texts [7]. Knowledge is implicitly stored in the weights of these models, which often contain billions of parameters, enabling them to have certain knowledge on open-domain topics [25]. Unfortunately, even the largest dialogue models suffer from the hallucination of knowledge, which can be interpreted as a form of lossy compression when employing training to encode that knowledge within the weights of a neural network [13]. The phenomenon of knowledge hallucination is prevalent in knowledge-intensive dialogue, which can significantly erode the user's trust in the model-generated results.

This study was supported by the National Natural Science Foundation of China under Grant 51978300.

F. Liu et al. (Eds.): NLPCC 2023, LNAI 14302, pp. 16–28, 2023.
https://doi.org/10.1007/978-3-031-44693-1_2

Recent works, such as RAG [16] and FiD [9], have augmented model generation by integrating external knowledge sources, known as retrieval-augmented text generation. This approach effectively mitigates the problem of knowledge hallucination. Retrieval-augmented text generation is a new text generation paradigm that fuses emerging deep learning technology and retrieval technology [17]. Compared with generation-based counterpart, this new paradigm not only improves the correctness of the model to generate factual answers, but also helps avoid safe but boring responses by leveraging external knowledge that is not present in the dialogue history. Besides, without significantly affecting model performance, the knowledge is not necessary to be implicitly stored in model parameters, but is explicitly acquired in a plug-and-play manner, leading to great scalability [17]. For example, Retro [1] obtains comparable performance to GPT-3 and Jurassic-1 [18], despite using 25× fewer parameters.

The key to improving the quality of generation is how to extract useful information from retrieved external documents and integrate them into answers. Specifically, on the dialogue response generation task, exemplar/template retrieval as an intermediate step has been shown beneficial to informative response generation [3,29,30]. Therefore, our work proposes a demonstration-based learning method, as shown in Fig. 1. We concatenate context and document with demonstration examples retrieved from the training set and feeds them into the generator. Subsequently, only the document and context segments generated by the encoder are utilized as input for the decoder to compute multi-head attention. The goal is to provide the generator with a better understanding of the task from the demonstrations so that it can make better use of the retrieved documents. As a plug-and-play module, our approach can be applied to various existing encoder-decoder models. To our knowledge, our work is the first to use demonstration-based learning to improve generator structures in knowledge-intensive dialogue.

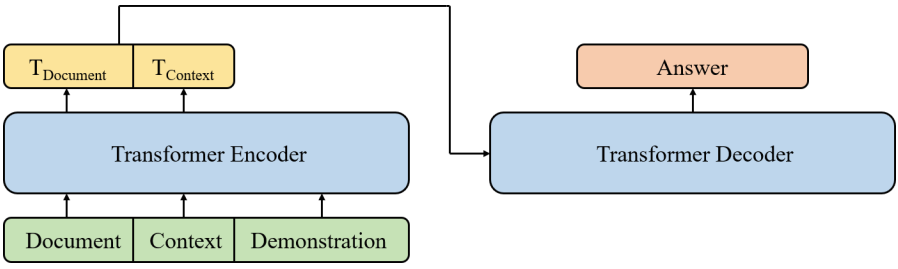

Fig. 1. Demonstration-based learning framework for generator

In knowledge-intensive dialogue, the accuracy of the retrieval model plays a crucial role in generating subsequent responses. [20] demonstrated for the first time from both theoretical and experimental perspectives that when the document is longer, it is difficult for a single fixed-dimensional vector to effectively represent the semantic information of the document. To obtain better results, it is necessary to expand the dimension of the representation vector or increase

the number of representation vectors. In this paper, we propose **RM-BERT**, a representation-interaction ranking model based on **r**esidual **m**ulti-vector over **BERT**, as shown in Fig. 2. By increasing the number of representation vectors, RM-BERT can better represent the semantic information of both documents and contexts, thereby improving retrieval performance. While conceptually simple, RM-BERT outperforms strong baselines on several benchmarks and achieves similar performance to BERT, while computing much faster than BERT.

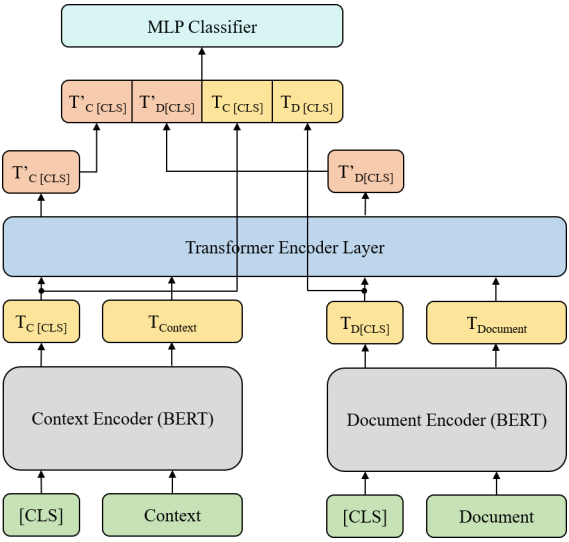

Fig. 2. Architecture of the RM-BERT

2 Related Work

Dense Retriever. According to the different ways of encoding the context and document as well as of scoring their similarity, dense retrievers can be roughly divided into three types [31]. Representation-based models, such as Realm [8] and DPR [11], are known for their speed, as the representations of documents can be computed and indexed offline in advance. However, the accuracy of these models is often compromised, as they only capture shallow interactions. Interaction-based models, such as BERT [5], can offer high accuracy by allowing for deep interactions between context and document. However, the heavy computation required by these models is not practical for real-world applications. To strike a balance between effectiveness and efficiency, a representation-interaction model is preferred. Many of these models use a late interaction architecture, such as ColBERT [12]. This architecture can not only compute the document representation in advance but also allow deep interactions between context and document, thus offering a good trade-off between speed and accuracy.

Generator. Recent research has focused on how to train a generator to better use retrieved external knowledge to improve the quality of the generated output. In previous studies, most models such as RAG and FiD just concatenate document and context and then input them into the model. They hope that the model will directly learn how to use the retrieved documents during the gradient descent. In contrast to these approaches, our proposal uses demonstration-based learning. In the supervised learning setting, the text most similar in distribution to the data in inference is the training data [28]. Therefore, we retrieve data from the training set as demonstration examples. By providing appropriate training data as demonstration examples, the model can gain a better understanding of the task and produce improved answers.

3 Methodology

Consider a conditional generation task where the input is a context c and the answer y is a sequence of tokens. To achieve better generation results in knowledge-intensive dialogue, external documents $D = \{d_1, d_2, ..., d_n\}$ are usually introduced. Retrieval-augmented text generation is the approach of adding external documents D for the model to condition on during its generation of y.

First, the retriever retrieves the top-k documents $D_k = \{d_1, d_2, ..., d_k\}$ most relevant to context c. Retriever can either directly retrieve the entire documents, or perform a preliminary retrieval first, and then sort the results. The generator then takes the context c and documents D_k as input to predict the answer y. Our generator is an encoder-decoder architecture designed according to the sequence-to-sequence modeling paradigm [26]. Specifically, the generator is based on the Transformer [27] architecture (e.g., BART [15] and T5 [23]) and parametrized by θ. We use the generator to model $p(y \mid c, D_k; \theta)$, where c and D_k are encoded by the bidirectional encoder, and the decoder predicts y autoregressively (conditioned on the encoded c and D_k and its left context). The likelihood of $p(y \mid c, D_k; \theta)$ is defined as:

$$p\left(y \mid c, D_k; \theta\right) = \prod_{t=1}^{N} p\left(y^t \mid y^{1:t-1}, c, D_k; \theta\right) \tag{1}$$

where N is the number of answer tokens.

Next, we introduce our generator improvement method and ranking model RM-BERT.

3.1 Generator

An illustration of our demonstration-based learning framework for generator is shown in Fig. 1. In the traditional retrieval-augmented generation, only document and context are input into the encoder. As an improvement we add additional demonstration examples as input to the encoder. In contrast to existing approaches that require additional human effort to generate such auxiliary

supervisions, our demonstrations can be automatically constructed by retrieving appropriate data from the training set. More precisely, we use BM25 [24] to retrieve training instances from the training set. Following the encoding and semantic interaction of the above three by the encoder, only the document and context portions produced by the encoder are utilized as input for the decoder to compute multi-head attention.

For a given training set $T = \{(c_1, y_1), ..., (c_m, y_m)\}$, we index it into a list of key-value pairs, where c_i is the context and y_i is the ground-truth label. Given a context c_i, we search for the top-q most similar contexts in the index as demonstrations G_q. Note that during training, as the context c_i is already indexed, we filter it from the retrieval results to avoid data leakage. Then we concatenate demonstrations with context c_i and retrieved external documents $D_k = \{d_1, d_2, ..., d_k\}$ to feed into the encoder. In the multi-turn dialogue scene, we only use the answer as demonstration, so $G_q = \{y_{i_1}, ..., y_{i_q}\}$ and the input form is $[d_j; c_i; y_{i_1}; ...; y_{i_q}]$. For the single-turn dialogue scene similar to question answering, we concatenate the context and answer as demonstration, so $G_q = \{(c_{i_1}, y_{i_1}), ..., (c_{i_q}, y_{i_q})\}$ and its input form is $[d_j; c_i; c_{i_1}; y_{i_1}; ...; c_{i_q}; y_{i_q}]$. d_j is one of the documents of D_k, so the actual input is a batch of size k. After adding the demonstration examples, Eq. 1 is rewritten as:

$$p\left(y_i \mid c_i, D_k, G_q; \theta\right) = \prod_{t=1}^{N} p\left(y_i^t \mid y_i^{1:t-1}, c_i, D_k, G_q; \theta\right) \tag{2}$$

where N is the number of answer tokens. The final loss is defined as the negative log-likelihood:

$$L_{NLL} = -\log p\left(y_i \mid c_i, D_k, G_q; \theta\right) \tag{3}$$

During inference, we will not filter any retrieved information, as all the retrieve data only come from training set.

3.2 RM-BERT

As Fig. 2 illustrates, the lower portion of RM-BERT is a representation-based model similar to DPR. However, unlike DPR, the two encoders we employ share parameters. Then we extend the representation-based model by adding a Transformer encoder layer. The motivation for this is that [21] has shown that the lower biLM layers specialize in local syntactic relationships, allowing the higher layers to model longer range relationships such as coreference, and to specialize for the language modeling task at the top most layers. To ensure retrieval efficiency, we only use one layer of Transformer encoder. Finally, we add a MLP classifier behind the Transformer encoder layer for classification.

During training, we first use two encoders to capture the local syntactic relationships of context and document respectively. Then we concatenate the outputs of these two encoders and feed them into the subsequent Transformer encoder layer to model longer range relationships. To represent semantic information more fully, we concatenate $T_{C[CLS]}$ and $T_{D[CLS]}$ output by the context encoder and document encoder with $T'_{C[CLS]}$ and $T'_{D[CLS]}$ output by the

Transformer encoder layer. Finally, they are fed into the MLP classifier for classification. $T_{C[CLS]}$ and $T_{D[CLS]}$ focus on each individual sentence and play a role in expanding semantic information. $T^{'}_{C[CLS]}$ and $T^{'}_{D[CLS]}$ focus on the semantic information after interaction and play a role in fully understanding the relationship between context and document. Given a collection of contexts $C = \{c_1, c_2, ..., c_m\}$ and a collection of documents $D = \{d_1, d_2, ..., d_n\}$, the relevance score of c_i to d_j, denoted as $s(c_i, d_j)$, is estimated via a MLP classifier:

$$s\left(c_i, d_j\right) = MLP\left(Concat\left(T^{'}_{C[CLS]}, T^{'}_{D[CLS]}, T_{C[CLS]}, T_{D[CLS]}\right)\right) \tag{4}$$

The probability of a document d_j being relevant to the context c_i is calculated as:

$$p\left(d_j \mid c_i, D\right) = \frac{exp\left(s\left(c_i, d_j\right)\right)}{\sum_{k=1}^{|D|} exp\left(s\left(c_i, d_k\right)\right)} \tag{5}$$

The ranking model parameters are updated by minimizing the cross-entropy loss:

$$L = -\sum_{(c_i, d_j)} \left(z_{i,j} log\left(p_{i,j}\right) + \left(1 - z_{i,j}\right) log\left(1 - p_{i,j}\right)\right) \tag{6}$$

where $z_{i,j}$ is the ground-truth label of c_i and d_j and $p_{i,j}$ is equivalent to $p(d_j \mid c_i, D)$.

During inference time, we first apply the document encoder to all the documents and save the encoding results offline. At runtime, only the context needs to be fed into the context encoder for encoding. The context encoding result is then fed into the Transformer encoder layer together with the precomputed document encoding result.

4 Experiments

In this section, we will introduce more details about experiments and the corresponding analysis.

4.1 Datasets

We conduct experiments on three datasets: Wizard of Wikipedia [6], Natural Questions [14] and TriviaQA [10]. Wizard of Wikipedia is a large dataset of

Table 1. Dataset statistics

Task	Dataset	Train	Dev
multi-turn dialogue	WoW	63734	3054
single-turn dialogue	NQ	87372	2837
single-turn dialogue	TriviaQA	61844	5359

multi-turn dialogue grounded with knowledge retrieved from Wikipedia. The input is a short dialog history ending with the information seekers turn. Natural Questions and TriviaQA are question answering datasets, which are equivalent to single-turn dialogue scenario in knowledge-intensive dialogues. Unlike the original versions, the relevant Wikipedia page must be found by a retrieval step. Overall statistics can be found in Table 1.

To retrieve the necessary information, we employ the standard KILT Wikipedia dump[1] [22]. Without loss of generality, we only merge the ground-truth label documents of the three datasets as our knowledge source for retrieval.

4.2 Evaluation Metrics

We employ standard KILT automatic metrics. KILT contains multiple evaluation metrics, which can be roughly divided into three types: (1) downstream results, (2) performance in retrieving relevant evidence to corroborate a prediction and (3) a combination of the two [22].

For retrieval task, in order to measure the correctness of the provenance, we adopt R-Precision and Recall@5. R-precision, calculated as r/R, where R is the number of Wikipedia pages inside each provenance set and r is the number of relevant pages among the top-R retrieved pages. Recall@k, calculated as w/n, where n is the number of distinct provenance sets for a given input and w is the number of complete provenance sets among the top-k retrieved pages [22].

For generation task, Wizard of Wikipedia uses Rouge-L, F1, KILT-RL and KILT-F1 to measure the correctness of the generated output. Natural Questions and TriviaQA use EM, F1, KILT-EM, KILT-F1. The KILT scores only award EM, ROUGE-L and F1 points to KILT-EM, KILT-RL and KILT-F1 respectively, if the R-precision is 1 [22]. This metric is employed to emphasize the importance of systems being able to substantiate their output with appropriate evidence, rather than just providing an answer.

4.3 Implementation Details

The BM25 library we use is based on Anserini[2]. Our models training is based on Transformers library[3]. All models use the AdamW optimizer [19] and fine-tune on each dataset independently after loading the pre-trained weights.

Retrieval. BM25 and DPR directly retrieve from the entire knowledge source, while ColBERT, BERT and our RM-BERT act as rerankers to rerank the top 100 results retrieved by BM25. For DPR and ColBERT, we use the same setting as the original papers. For BERT, we employ the BERT base model. Both encoders for RM-BERT are also BERT base models, and they share parameters. The transformer encoder layer of RM-BERT is initialized using the top layer weights

[1] https://github.com/facebookresearch/KILT.
[2] https://github.com/castorini/anserini.
[3] https://github.com/huggingface/transformers.

of pre-trained BERT. For the MLP classifier, we use two linear layers with a ReLU activation function for nonlinear transformation.

Generation. To show the efficacy of demonstration-based learning as a plug-and-play method, we present performance in two models: RAG and FiD. Specifically, we use the RAG-Token model, and for FiD we use FiD-base. Referring to the original paper's settings, RAG uses 5 retrieved documents, and FiD uses 20 retrieved documents. For the Wizard of Wikipedia dataset, we use one demonstration, while for the Natural Questions and TriviaQA datasets, we use three demonstrations. We also consider demonstrations of other quantities and compare them empirically in Sect. 4.5. In addition, we also select BART and T5 as baselines, two models that do not retrieve knowledge sources, and their experimental results are sourced from [22].

4.4 Results

Table 2. Results for Retrieval

Model	WoW		NQ		TriviaQA	
	R-Prec	Recall@5	R-Prec	Recall@5	R-Prec	Recall@5
Bm25	37.07	61.23	49.10	65.04	51.37	69.44
DPR	40.86	65.62	61.02	69.69	60.94	68.28
ColBERT	44.47	70.37	65.84	76.56	68.43	79.42
RM-BERT	46.30	71.25	67.04	77.12	68.67	79.01
BERT	47.51	72.43	68.77	78.21	71.21	81.41

Retrieval. Table 2 illustrates the retrieval evaluation results of our proposed model in comparison with the baselines. The results indicate that our model achieves performance close to BERT across all datasets, while requiring considerably less computational resources. Furthermore, in comparison with ColBERT, which is also a representation-interaction model, our approach exhibits superior performance on nearly all datasets. Despite BM25 and DPR's advantage in terms of speed, they significantly lag behind our model concerning retrieval effectiveness. Overall, our model achieves an optimal trade-off between efficacy and efficiency.

Generation. In Tables 3, 4 and 5, we present the performance evaluation results of our proposed approach and the baselines. Firstly, we observe that the inclusion of the demonstration-based learning approach leads to substantial improvements in nearly all evaluation metrics across all datasets for both RAG and

Table 3. The results on Wizard of Wikipedia dataset

Model	Rouge-L	F1	KILT-RL	KILT-F1
BART	12.05	13.35	0	0
T5	12.80	13.28	0	0
RAG	16.03	18.10	10.61	12.00
FiD	14.81	16.73	9.79	11.07
RAG-demo	16.64(+0.61)	18.87(+0.77)	10.71(+0.10)	12.20(+0.20)
FiD-demo	15.70(+0.89)	17.56(+0.83)	9.85(+0.06)	11.13(+0.06)

Table 4. The results on Natural Questions dataset

Model	EM	F1	KILT-EM	KILT-F1
BART	26.15	32.06	0	0
T5	25.20	31.88	0	0
RAG	46.14	53.94	40.25	45.46
FiD	48.78	57.00	38.53	44.47
RAG-demo	47.83(+1.69)	55.41(+1.47)	41.63(+1.38)	46.83(+1.37)
FiD-demo	51.50(+2.72)	59.68(+2.68)	41.17(+2.64)	46.98(+2.51)

Table 5. The results on TriviaQA dataset

Model	EM	F1	KILT-EM	KILT-F1
BART	32.54	39.58	0	0
T5	25.79	33.72	0	0
RAG	49.60	61.92	40.57	48.80
FiD	51.26	65.87	39.39	49.50
RAG-demo	53.41(+3.81)	64.69(+2.77)	42.30(+1.73)	49.60(+0.80)
FiD-demo	54.75(+3.49)	64.23(-1.64)	42.02(+2.63)	48.52(−0.98)

FiD models, as shown in the corresponding improvements indicated in parentheses. This finding not only confirms the efficacy of our proposed approach but also demonstrates its applicability to a wide range of models. Additionally, the results indicate the usefulness of our method for both single-turn and multi-turn knowledge-intensive dialogues. Next, we compare the performance of models with and without the usage of knowledge sources. We observe a significant improvement in model performance after incorporating knowledge sources, as evidenced by the comparison results. This highlights the critical role of knowledge retrieval in knowledge-intensive dialogues.

4.5 Ablations

To explore the effect of RM-BERT, we conduct relative ablation studies, as illustrated in Table 6. The RM-BERT-single model feeds only a single vector $T'_{C[CLS]}$ into the MLP for classification. Our findings suggest that reducing the number of representation vectors resulted in decreased model performance across all datasets, with the most significant impact observed on the R-Prec metric for the Wizard of Wikipedia dataset. These results underscore the inadequacy of using only one vector to represent the semantic information of sentences and demonstrate the efficacy of our proposed method in addressing this limitation.

We also conduct ablation studies to investigate the impact of the number of demonstrations on the model's performance. The results of these studies are presented in Table 7. Our findings indicate that, in the case of multi-turn dialogue, the model's performance is optimal when only one demonstration is used. As the number of demonstrations increases, the model's performance experiences a slight decrease. Conversely, for single-turn dialogue, the model's performance is best when three demonstration examples are used, with any additional demonstration examples leading to a degradation in performance.

Table 6. Ablation experiment for the variation of RM-BERT

Model	WoW		NQ		TriviaQA	
	R-Prec	Recall@5	R-Prec	Recall@5	R-Prec	Recall@5
RM-BERT-single	40.47	71.09	66.13	76.81	65.07	78.32
RM-BERT	46.30	71.25	67.04	77.12	68.67	79.01

Table 7. The effect of different number of demonstrations on the performance of FiD-demo model

Demo num	WoW				NQ			
	Rouge-L	F1	KILT-RL	KILT-F1	EM	F1	KILT-EM	KILT-F1
1	15.70	17.56	9.85	11.13	49.24	57.64	39.27	45.31
2	15.48	17.15	9.71	10.84	50.90	59.08	40.54	46.41
3	15.12	16.69	9.30	10.36	51.50	59.68	41.17	46.98
4	15.14	16.61	9.27	10.31	51.22	59.50	41.03	46.82
5	15.00	16.41	9.10	10.09	51.22	59.37	40.89	46.56

4.6 Case Study

We conduct a case study on the WoW and NQ dev sets to intuitively compare our model with the baseline, as presented in Table 8. On the WoW task, FiD

generates incorrect answer due to insufficient understanding of the context, while on the NQ task, it directly generates factually incorrect answer. In contrast, FiD-demo generates more specific and factually accurate responses.

Table 8. Case study on the WoW and NQ dev sets

Task	Input	Model	Output
WoW	User: My favorite color is red do you like it? Bot: I like red, but pink is my favorite it is named after a flowering plant in the genus Dianthus. User: What is the wavelength of red?	FiD	It was first used as a color name in the late 17th century.
		FiD-demo	I'm not sure, but I know it is the color at the end of the visible spectrum of light, next to orange and opposite violet.
NQ	Who has made the most premier league appearances?	FiD	Gary Speed
		FiD-demo	Gareth Barry

5 Conclusions

In this paper, we present a new method to improve the generator in knowledge-intensive dialogue and propose a ranking model RM-BERT. We propose to improve the generator through demonstration-based learning, allowing the generator to better utilize the retrieved knowledge sources by enhancing model's understanding of the task. Experiments prove that, as a plug-and-play method, it is applicable not only to multiple datasets, but also to a variety of models. Furthermore, we use multiple vectors to better represent the semantic information of context and document in a way similar to residual connections. Despite having a much lower computational cost than BERT, RM-BERT achieves a performance very close to BERT, which is a significant achievement in practical application scenarios. In future work, we intend to explore combining our demonstration-based learning method with other methods, potentially providing orthogonal improvement. We also intend to further explore how to represent sentences more effectively and efficiently.

References

1. Borgeaud, S., et al.: Improving language models by retrieving from trillions of tokens. In: International Conference on Machine Learning, pp. 2206–2240. PMLR (2022)
2. Brown, T., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
3. Cai, D., et al.: Skeleton-to-response: dialogue generation guided by retrieval memory. arXiv preprint arXiv:1809.05296 (2018)
4. Chowdhery, A., et al.: Palm: scaling language modeling with pathways. arXiv preprint arXiv:2204.02311 (2022)

5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., Weston, J.: Wizard of wikipedia: knowledge-powered conversational agents. arXiv preprint arXiv:1811.01241 (2018)
7. Glass, M., Rossiello, G., Chowdhury, M.F.M., Naik, A.R., Cai, P., Gliozzo, A.: Re2g: retrieve, rerank, generate. arXiv preprint arXiv:2207.06300 (2022)
8. Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.: Retrieval augmented language model pre-training. In: International Conference on Machine Learning, pp. 3929–3938. PMLR (2020)
9. Izacard, G., Grave, E.: Leveraging passage retrieval with generative models for open domain question answering. arXiv preprint arXiv:2007.01282 (2020)
10. Joshi, M., Choi, E., Weld, D.S., Zettlemoyer, L.: TriviaQA: a large scale distantly supervised challenge dataset for reading comprehension. arXiv preprint arXiv:1705.03551 (2017)
11. Karpukhin, V., et al.: Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2004.04906 (2020)
12. Khattab, O., Zaharia, M.: ColBERT: efficient and effective passage search via contextualized late interaction over BERT. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 39–48 (2020)
13. Komeili, M., Shuster, K., Weston, J.: Internet-augmented dialogue generation. arXiv preprint arXiv:2107.07566 (2021)
14. Kwiatkowski, T., et al.: Natural questions: a benchmark for question answering research. Trans. Assoc. Comput. Linguist. **7**, 453–466 (2019)
15. Lewis, M., et al.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461 (2019)
16. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv. Neural. Inf. Process. Syst. **33**, 9459–9474 (2020)
17. Li, H., Su, Y., Cai, D., Wang, Y., Liu, L.: A survey on retrieval-augmented text generation. arXiv preprint arXiv:2202.01110 (2022)
18. Lieber, O., Sharir, O., Lenz, B., Shoham, Y.: Jurassic-1: technical details and evaluation. White Paper. AI21 Labs 1 (2021)
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
20. Luan, Y., Eisenstein, J., Toutanova, K., Collins, M.: Sparse, dense, and attentional representations for text retrieval. Trans. Assoc. Comput. Linguist. **9**, 329–345 (2021)
21. Peters, M.E., Neumann, M., Zettlemoyer, L., Yih, W.T.: Dissecting contextual word embeddings: architecture and representation. arXiv preprint arXiv:1808.08949 (2018)
22. Petroni, F., et al.: KILT: a benchmark for knowledge intensive language tasks. arXiv preprint arXiv:2009.02252 (2020)
23. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(1), 5485–5551 (2020)
24. Robertson, S., Zaragoza, H., et al.: The probabilistic relevance framework: Bm25 and beyond. Found. Trends® Inf. Retr. **3**(4), 333–389 (2009)
25. Shuster, K., Poff, S., Chen, M., Kiela, D., Weston, J.: Retrieval augmentation reduces hallucination in conversation. arXiv preprint arXiv:2104.07567 (2021)

26. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, vol. 27 (2014)
27. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
28. Wang, S., et al.: Training data is more valuable than you think: a simple and effective method by retrieving from training data. arXiv preprint arXiv:2203.08773 (2022)
29. Weston, J., Dinan, E., Miller, A.H.: Retrieve and refine: improved sequence generation models for dialogue. arXiv preprint arXiv:1808.04776 (2018)
30. Wu, Y., Wei, F., Huang, S., Wang, Y., Li, Z., Zhou, M.: Response generation by context-aware prototype editing. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 7281–7288 (2019)
31. Zhu, F., Lei, W., Wang, C., Zheng, J., Poria, S., Chua, T.S.: Retrieving and reading: a comprehensive survey on open-domain question answering. arXiv preprint arXiv:2101.00774 (2021)