Title: Query Expansion with Topic-aware In-context Learning and Vocabulary Projection for Open-domain Dense Retrieval

Authors:
1. Ronghan Li, School of Computer Science and Technology, Xidian University
266 Xifeng Road, Chang'an District, Xi'an, PR China, 710126, lironghan@xidian.edu.cn, +8618792632084
2. Mingze Cui, College of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning, PR China, 116023, mingze_cui@foxmail.com, +8615619246908
3. Benben Wang, School of Computer Science and Technology, Xidian University
266 Xifeng Road, Chang'an District, Xi'an, PR China, 710126, 22009200148@stu.xidian.edu.cn, +8617639162276
4. Yu Wang, School of Computer Science and Technology, Xidian University
266 Xifeng Road, Chang'an District, Xi'an, PR China, 710126, FranciscoSmith448@gmail.com, +8613125864594
5. Qiguang Miao, School of Computer Science and Technology, Xidian University (Corresponding Author, qgmiao@xidian.edu.cn)，266 Xifeng Road, Chang'an District, Xi'an, PR China, 710126，+8618681880036

Query Expansion with Topic-aware In-context Learning and Vocabulary Projection for Open-domain Dense Retrieval

Ronghan Li[a], Mingze Cui[b], Benben Wang[a], Yu Wang[a], Qiguang Miao[a,*]

[a]*School of Computer Science and Technology, Xidian University, Xi'an, 710126, Shaanxi, China*
[b]*College of Computer Science and Technology, College of Computer Science and Technology, Dalian, 116023, Liaoning, China*

Abstract

Large language models (LLMs) have recently emerged as pivotal components in open-domain question answering. This study proposes a simple yet effective method to enhance dense retrieval using topic-aware In-Context Learning (ICL) and topic keyword projection. First we leverage LLMs to generate a pseudo-passage based on topic-aware demonstrations obtained from the pre-trained cluster to which the target query belongs. Second, we employ the masked language model (MLM) header in the autoencoder LM to map the query representation to implicitly topic-related tokens as keywords. We combine these two approaches to augment the original query. Extensive experiments on four prevalent open-domain question answering (ODQA) datasets demonstrate that our method achieves an average of 4.26% improvement in R@20 compared to state-of-the-art query expansion work. Further analysis shows that the relevant demonstrations can provide higher-quality pseudo-passage generation, and the extracted keywords provide an interpretable basis for the effectiveness of dense retrieval. Code and data are available at `https://github.com/XD-BDIV-NLP/TDPR`.

Keywords: Query Expansion, Open-domain Retrieval, In-context Learning

*Corresponding author
Email address: `qgmiao@xidian.edu.cn` (Qiguang Miao)

1. Introduction

Retrieval-augmented generation (RAG) is used to tackle complex natural language processing tasks that require extensive knowledge [2]. RAG combines the strengths of retrieval and generation by utilizing a retrieval model to extract pertinent information from a knowledge base and integrating this information with a generation model to create more informative and relevant responses. In recent years, there have been improvements and studies on RAG, including enhancing the model's ability to generate relevant and accurate responses [3], increasing its adaptability in specific domains [4], and incorporating auxiliary rationale memories to improve the problem-solving abilities of large language models, thus enhances the model's capability to generate more accurate rationales [5]. Dense passage retrieval (DPR), which projects text and queries into low-dimensional vector representations, has shown remarkable effectiveness [6, 7]. DPR typically employs two separate encoders to obtain the passage and query representations, followed by similarity matching methods like dot product or cosine similarity. Based on the DPR framework, significant efforts have been devoted in the literature to improve the representation capabilities of encoders during the pretraining and finetuning stages. For the former, strategies such as introducing a transformer encoder to enhance the representation capability of the `[CLS]` vector [8], or introducing contrastive learning tasks to optimize the representation space [9]. As for the finetuning stage, data augmentation and negative sample mining are generally considered effective promotion methods. For negative sample sampling, a strategy involves continuously re-encoding the entire document library based on the updated DPR model during training [10]. Data augmentation, on the other hand, involves improving DPR by leveraging a pretrained generative model (e.g., T5) to generate queries for each paragraph, thereby constructing query-passage training data pairs [11]. Although adding more query-passage pairs has proven effective, supplementing existing data with extra information has not been thoroughly explored.

Query expansion [12, 13] is a commonly employed technique in information retrieval, aiming to supplement information omitted from natural language queries by adding related terms or sentences, thereby increasing the accuracy and comprehen-

siveness of retrieval. By utilizing advanced language models, optimizing retrieval metrics [14], the relevance feedback mechanism in information retrieval systems can be enhanced, thereby achieving more accurate document retrieval. Recently, LLMs such as ChatGPT [15] and LLaMA [16] have made significant strides in information retrieval. For example, taking gold query-passage training samples as in-context demonstrations, Query2doc [17] expand the query by utilizing LLMs to generate pseudo-passages related to the query. Nevertheless, since the demonstrations are randomly sampled, ensuring their topical relevance to the current query for expansion becomes challenging, which may mislead LLMs to generate irrelevant pseudo-passages or even factuality hallucination. It has been documented that relevant examples and selection strategies help large models better understand the test problem and thus generate higher-quality content [18]. For example, as shown in Fig. 1, although the demonstrations in the left sub-graph are not semantically ***relevant*** to the test question, the topic of asking "who" is enough for the LLM to generate a pseudo-passage containing the answer. In contrast, the randomly selected demonstrations only provide ***related but insufficient*** information for the LLM.

Therefore, open-domain dense retrieval research mainly faces the following three challenges:

- *How to unify and integrate various data features for dense retrieval tasks?* In many retrieval systems, especially those utilizing large language models, different aspects of the data (such as text semantics, query structure, and latent representations) need to be integrated effectively to enhance retrieval performance. Our method focuses on seamlessly integrating these multi-faceted features, aiming to provide more relevant search results in response to user queries.
- *How to accurately extract the essential information from queries and documents for dense retrieval?* Retrieving the most relevant information requires accurately identifying and representing key aspects of both queries and documents. Our work aims to improve this process by leveraging topic-aware context learning, which aids in better pseudo-passage generation and improves the retrieved information's relevance.

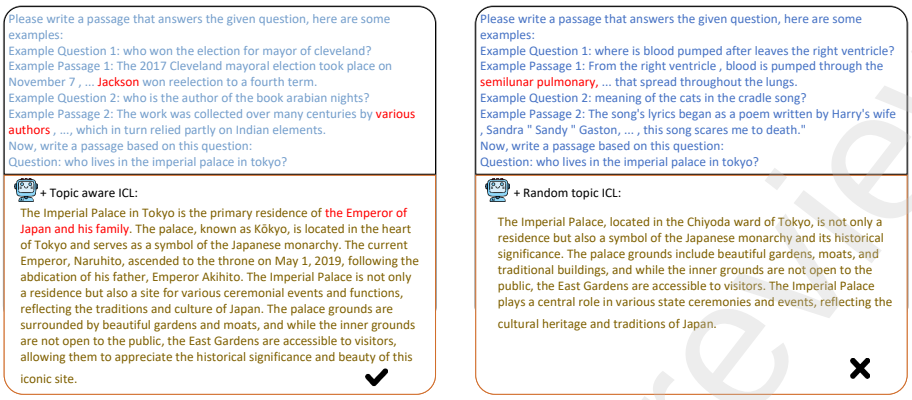


Fig. 1. An illustration of topic-aware demonstration. Topic-relevant demonstrations help the LLM generate better pseudo-passage (left) while related demonstrations guide the LLM to generate the pseudo-passage with insufficient information.

- *How to reduce computational overhead while enhancing retrieval performance?* Many state-of-the-art retrieval methods face challenges regarding computational efficiency. Our approach, TDPR, addresses this by directly incorporating key information extracted by a masked language model into the original query, thus reducing the computational complexity while maintaining high retrieval accuracy.

To alleviate this problem, in this work, we introduce topic-aware in-context learning to improve query expansion. Specifically, different from existing retrieval-based and posterior-based methods [19, 20], we first cluster the training queries via unsupervised methods (e.g., k-means), which can be viewed as constructing several topic pools based on the aspects on which the queries focus. Then, we sample in-context demonstrations for the current query from the cluster to which it belongs. With these topic-related demonstrations, we believe LLMs can generate more relevant pseudo-passages and reduce factual hallucinations. Additionally, existing work [1] has shown that DPR query encoders can implicitly reflect related tokens on vocabulary projections. Inspired by this intuition, we further expand the query with relevant projection tokens that do not appear in the query and its pseudo-passage.

With the expanded queries, we retrain DPR, and our experiments on four prevalent ODQA datasets demonstrate that employing topic-aware in-context learning and relevant projection tokens to augment query context improves performance ranging from 3.5% to 7.4% compared to the vanilla DPR [6], obtaining competitive results against other state-of-the-art dense retrieval baselines. We further performed ablation experiments to confirm the effectiveness of the proposed method. Additionally, since the method proposed in this paper focuses on augmenting queries, it can be used in conjunction with other methods. The main contributions of this paper are as follows:

- We present a new ICL method for dense retrieval, employing a topic-aware demonstrations to enhance the quality of generated pseudo-passages.
- To further improve retrieval performance via potentially relevant topics, we combine projection tokens relevant to the original query with pseudo-passages obtained through ICL.
- We conduct comprehensive experiments to verify the effectiveness of the proposed method and examine existing challenges as well as future research directions.

2. Related Work

2.1. Open-Domain Question Answering

ODQA [21] seeks to leverage extensive datasets like Wikipedia [22] or BookCorpus [23] to answer factoid querys. A classic ODQA system usually contain an information retriever and a reader. While the target of reader like BERT [24] or T5 [25] is to understanding and reasoning the retrieved evidences and yield the answer, the information retrieverfocuses on extracting text relevant to the query from a vast knowledge base. Dense Passage Retriever(DPR) [6] is one of the popularly used IR in ODQA. Despite DPR have shown great success, there is a lack of comprehensive exploration into interpreting their representations.Analyzing neural retrieval models and utilize diagnostic probes [26]tried to test characteristics such as sensitivity to paraphrases, stylistic variations, and factual accuracy, extending their study beyond dense

retrievers. Decoding the query representations of neural retrievers using T5 decoder [27] can get better queries for retrieval, while using BERT mlm head [1] show there are hidden information within the representations and can be use for lexical enrichment. Condenser [8] introduced a two-layer transformer encoder in DPR training to aggregate text information into the dense representation, by constructing negative samples from Approximate Nearest Neighbor (ANN), ANCE [10]updates them throughout the learning process. Topic-DPR [28] leverages contrastive learning to simultaneously optimize multiple topic-based prompts, enhancing the uniformity of text representations in dense retrieval. These prompts are defined on a probabilistic simplex, aiding LLMs in better understanding and processing documents related to specific topics, thereby improving semantic representations of documents, and optimized through contrastive learning. Compares to our approach, Topic-DPR focuses more on exploring the complex relationships between topics through multiple topic-based prompts and integrating these relationships into PLMs to enhance retrieval performance. GENREAD [29] utilizes the generative capabilities of large pre-trained language models (such as InstructGPT) to directly generate contextual documents relevant to a given query. Additionally, it employs a clustering-based prompting method, where different query-document pairs sampled from clusters generate documents that cover various perspectives of the query, thereby enhancing the coverage and performance of answers. This method focuses more on the quality of documents generated by LLMs, whereas the approach proposed in our paper is relatively less sensitive to the quality of LLM generation.

2.2. Query Expansion

Query expansion is one of the classical techniques in open-domain query answering. Generally, it's goal is to narrow down the lexical gap between querys and the documents. Query expansion often involves rewriting, such as rewriting the lexical resources [30], or rephrasing the query with input from relevance feedback [12, 13]. The pseudo-relevance feedback (PRF) framework, which integrates relevance matching with semantic matching [14], enhanced the quality of feedback documents through the use of the BERT model. Automatically generated query variants can be used to estimate document retrievability [31], thereby improving relevance feedback.

Representation-based ranking approach [26] uses a contextualized language model to model term importance, perform passage expansion, and ground representations in the lexicon, thereby improving retrieval effectiveness and reducing query time latency. BERT-QE [32] propose to conduct query expansion from the re-ranked passages and leverage overlapping chunks for next-phrase re-ranking. In situations where there are no labels provided, the top-k retrieved documents can be utilized as signals for pseudo-relevance feedback [33]. As LLMs become more popular, few-shot setting like Query2doc [17] randomly sampled k labeled pairs from training set as part of the in-context prompt. However, it focuses on sparse retrieval on MS MARCO and TREC DL. While prior research has explored optimizing query generation as an integrated module within RAG systems [34], subsequent investigations have revealed that expansion artifacts induced by LLM-generated content may propagate biases in dense retrievers [35, 36]. This challenge, where neural retrievers disproportionately favor machine-generated queries over human-formulated counterparts, has been identified as a critical methodological limitation [37]. Mitigating such model-induced distributional biases presents a promising direction for future research in developing robust RAG architectures.

2.3. In-context Learning with Selective Demonstrations

LLMs have demonstrated remarkable performance and potential recently. With billions of parameters, models like GPT-3 [38], and LLaMA [16] are trained on trillions of tokens, enabling them to handle tasks using either zero-shot approaches or in-context learning with few-shot prompts. HyDE [39], for instance, focuses on the zero-shot setting and uses pseudo-document embeddings for similarity-based retrieval.

In-context learning allows LLMs to observe a few examples within a context and apply the patterns they learn to solve a new problem. Research has shown that LLMs can successfully carry out complex tasks using in-context learning, such as solving mathematical reasoning problems [40]. Not only does this approach provide an interpretable interface for interacting with LLMs [38], it also makes it easier to incorporate human knowledge into LLMs by changing the demonstration and templates [18]. Several studies have investigated the impact of relevance and diversity in demonstrations

on LLM outputs [41]. Regarding demonstration selection strategies for in-context learning, KATE [18] use a kNN-based unsupervised retriever for selecting in-context examples.Alternatively, some methods build an unsupervised retriever (e.g., BM25) to recall similar examples and use a supervised retriever like EPR to choose the best demonstrations from these candidates [42]. Peng et al. [19] propose a data- and model-dependent demonstration selection method that identifies the best demonstrations by minimizing the cross entropy of the whole prompt and the demonstrations estimated by the inference model. However, these methods have not been validated on query expansion.

3. Background

In this work, we present a method to improve query expansion for dense passage retriever, by combining representation projection using language model head, and topic-aware in-context learning with LLM. To provide a comprehensive understanding, we commence with the essential background details.

3.1. DPR

Given a corpus of passages $C = \{p_1, \cdots, p_m\}$ and a query q, DPR utilizes bi-encoder to individually encode them, obtaining representations that can be used for subsequent relevance calculations between the query and the passages.

$$e_q = \mathrm{Enc}_Q(q) \tag{1}$$

$$e_p = \mathrm{Enc}_P(p) \tag{2}$$

Here, e_q denotes the encoding representation of the query, while e_p represents the encoding of the passages. Following the acquisition of the encoding representations for the query and passages, DPR calculates the similarity score by computing the dot product between the two representations.

$$\mathrm{sim}(q, p) = e_q^\top e_p \tag{3}$$

3.2. Query Expansion

Query expansion is a technique frequently employed in information retrieval and natural language processing to improve the effectiveness of search queries. The primary goal of query expansion is to broaden the scope of the original query, retrieve more relevant information, and improve the overall search results. This is accomplished by enriching the original query with additional terms or phrases that are semantically related.

Query expansion typically involves using a generative model (e.g `GPT-3.5`) to generate knowledge relevant to the given query:

$$q' = \text{Model}(q) \tag{4}$$

Here, q' denotes the expanded information generated by the model. In this way, knowledge from models other than DPR can be introduced to enhance the training of the DPR model.

4. Methods

This section introduces the methods used in this paper, structured as follows: Section 4.1 presents an overview. Section 4.2 introduces the ICL method used in this paper, which clusters queries in the dataset using k-means to give them a thematic sense. Section 4.3 describes how the MLM head is used to project query DPR vectors for feature extraction to obtain keywords. Section 4.4 details the fine-tuning of the retriever, where pseudo-passages and keywords are concatenated to the original queries as new queries. Section 4.5 explains the details of retrieval inference, where we employ the methods from Section 4.2 and Section 4.3 on the test set.

4.1. Overview

The overall process of this paper is shown in Fig. 2. We first employ the topic-aware ICL method, passing similar queries and passages into LLM through clustering to obtain higher-quality pseudo-passages. Simultaneously, we use the Topic Keywords

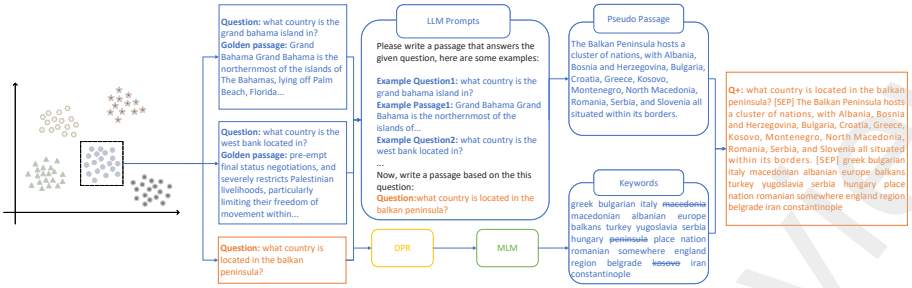


Fig. 2. TDPR uses clustering to categorize queries based on different topics. Two queries with the same topic and a ground-truth passage are jointly passed into the LLM to generate a pseudo-passage that can answer the target query. Simultaneously, the DPR vector of the target query is passed into the MLM head to obtain a series of interpretable keywords. Finally, the target query, pseudo-passage, and keywords are concatenated to form the final Q+. The strikethrough keywords represent those that have been deleted due to repetition with the query or the generated pseudo-passage.

Projection to obtain keywords related to the original query. These keywords will exclude stop words and words already present in the original query and pseudo-passage, aiming to acquire more information relevant to the original query but not existing in both. Finally, we concatenate the original query, pseudo-passage, and keywords to form a new query for subsequent training. Since DPR uses BERT-style models, we use `[SEP]` as the concatenation separator:

$$\mathbf{Q+} = \text{query [SEP] d+ [SEP] Keys} \tag{5}$$

4.2. Topic-Aware In-Context Learning

While performing feature projection, we employ LLM to generate a pseudo-passage. In this work, we do not utilize a zero-shot setting; instead, we propose a topic-aware approach for pseudo-passage generation. First, we input all query tokens into BERT and use the [CLS] hidden state d_i in the last layer as the representation of the query. Next, we perform the K-means algorithm on all the query representations. The K-means algorithm consists of two alternating steps: the assignment and update steps. These steps are performed iteratively until convergence. The assignment step can be expressed by the following equation:

$$C_{d_i}^t = \arg\min_j \|d_i - \mu_j^t\|_2 \tag{6}$$

In the equation, μ_j^t represents the j-th cluster's centroid when the algorithm executes at the t-th time step, and $C_{d_i}^t$ denotes the cluster closest to the i-th embedding under the Euclidean distance. After the assignment step, the algorithm updates the centroid of each cluster based on the cluster assignment of each embedding:

$$\mu_i^{t+1} = \frac{1}{|C_{d_i}^t|} \sum_{d_i \in C_{d_i}} d_i \tag{7}$$

Finally, we calculate the cluster nearest to the current query representation based on the Euclidean distance and randomly select two other queries that belong to that centroid:

$$\{x_a, x_b\} \in C_{d_i} = \arg\min_i \|d_i - \mu_i\|_2 \tag{8}$$

In this work, we select two random queries $\{x_a, x_b\}$ from the corresponding cluster for each query and input them into the LLM along with their golden passages and the target query as depicted in Figure 2. LLM, utilizing contextual learning, comprehends the topic of the query, provides a clear direction for answering, and generates pseudo-passage $d+$ related to the target query. The generated pseudo-passage $d+$ will be concatenated with the target query to create a new target problem for subsequent training of the DPR model.

4.3. Topic Keywords Projection

Within the fields of natural language processing and machine learning, representation projection is often employed to map the representation of text or other data from a specific embedding space to another, aiming better to capture relationships and semantic information between the data.

Representation projection involves learning a mapping function that projects input representations into target semantic space. In this work, we follow [1] to the Masked Language Model (MLM) head as a tool for representation projection. The role of the

MLM head is to take the hidden vectors $\boldsymbol{x} \in \mathbb{R}^d$ as input and return their distribution in the model's vocabulary space, defined as follows:

$$\mathbf{MLM-Head}(\boldsymbol{x})[i] = \frac{\exp(\boldsymbol{h}_i^\top) g(\boldsymbol{x})}{\sum_{j \in \mathcal{V}} \exp(\boldsymbol{h}_j^\top) g(\boldsymbol{x})} \tag{9}$$

Note that g is a non-linear function, and $\boldsymbol{h}_i \in \mathbb{R}^d$ corresponds to the word embedding of the i-th item in the vocabulary $\mathcal{V}$. Inspired by [1], we enhance the original query by integrating tokens derived from this mapping method as keywords to improve the quality of dense retrieval.

Given a query q, we first generate its DPR hidden representation h_q. Then, we feed it into an MLM (Masked Language Model) head to obtain the top-k topic-aware tokens $\mathbf{MLM-Head}(\boldsymbol{h_q})[1:k]$. Following [1], we directly use the MLM head of a BERT-style LM as the projection module. The query representation, after undergoing vocabulary projection, will yield top-k interpretable vocabulary tokens **Keys** related to the original query. We remove stop words from these words to reduce potential noise. Alternatively, we propose to exclude tokens existing in the original query q and pseudo-passage $d+$ from the top-k tokens, aiming to reduce explicit topic overlap. The process is formalized as follows:

$$\begin{aligned} \boldsymbol{h}_q &= \mathbf{DPR}(q) \\ \mathbf{Keys} &= \text{filter}(\mathbf{MLM-Head}(h_q)) \end{aligned} \tag{10}$$

where $filter$ denotes the function of the above operations. The resulting tokens will serve as keywords for subsequent query expansion in Eq. 5.

4.4. Retriever Finetuing

After combining the generated pseudo-passage and keywords obtained from the aforementioned process with the original query to form a new query, we retrain DPR using the official in-batch contrastive loss:

$$L(Q+, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-) = -\log \frac{e^{\text{sim}(Q+, p_i^+)}}{e^{\text{sim}(Q+, p_i^+)} + \sum_{j=1}^{n} e^{\text{sim}(Q+, p_{i,j}^-)}} \tag{11}$$

where $Q+$ represents the concatenated query, p_i^+ represents the gold passage, along with n irrelevant passage $p_{i,j}^-$. Each expanded question is paired with a gold passage, using the gold passages of other questions from the same mini-batch as negatives. Optionally, additional *hard* negatives can be considered, such as those retrieved by BM25 or the DPR checkpoint, which do not contain the answer but match most of the question tokens.

4.5. Inference

Following DPR [6], during the inference process, we similarly apply the passage encoder to all passages and index them using FAISS [43] offline. We also use the same method above on the test set, adding generated pseudo-passages and keywords to the original queries. In the test set, we measure the distance between the query vectors and the centroids obtained from the training set clustering. We randomly select two queries and ground-truth passages from the nearest centroid in the training set as examples for ICL. Given a query, we derive its embedding and return the top-k passages whose embeddings approximate it.

In actual applications, a similar process will be performed. Specifically, clusters are constructed in advance based on question-passage sets in general [44] or specific domains. Then, the user's query is used simultaneously for generating pseudo passages and projecting topic-aware tokens. Finally, the merge is fed into the trained retriever. We will analyze the time consumption in Section 6.5.

4.6. Discussion

In this section, we further elaborate on the theoretical and practical significance of the proposed method. Although existing methods have proven that relevant demonstrations are effective for high-quality generation, selective examples—particularly those that are topic-aware—have not been verified on dense retrieval. Compared with sparse retrieval, dense retrieval based on semantic matching requires more identification of the true intent, which is very important for queries containing omissions and references in practical scenarios. Therefore, if the large model can generate more accurate extended content based on topic-aware examples, it can improve the retrieval quality

in practical applications such as customer service and search engines and eliminate the need for clarifying words and additional retrieval input. Of course, the generation of large models based on contextual learning will bring higher latency (e.g., the time to generate pseudo-passages) and the hallucination problem, which we will discuss more in Section 6.5 and Section 6.6.

5. Experiment

5.1. Experiment Setup

5.1.1. Datasets

We have validated our work using four datasets. Below is a brief introduction to each dataset and its data details.

Natural Questions (NQ) [45] is tailored for end-to-end query answering tasks. The dataset comprises queries extracted from actual Google search queries, while the answers are specific spans within Wikipedia passages. This setup ensures the dataset mirrors real-world scenarios, making it valuable for training and evaluating query-answering systems.

TriviaQA (TQ) [46] is a dataset featuring a collection of trivia queries along with their corresponding answers. These queries and answers were initially obtained by scraping various sources on the web.

WebQuestions (WQ) [47] is a dataset of queries selected using the Google Suggest API. In this dataset, the answers correspond to entities found in Freebase.

CuratedTREC (TREC) [48] gathers queries from TREC QA tracks and various web sources. This dataset is specifically designed for open-domain query answering tasks using unstructured corpora.

5.1.2. Inplementation Details

We use `kmeans-pytorch`[1] to implement the query clustering, with Euclidean distance as the computation method. We set different cluster numbers based on the statistics of datasets. We cluster the training queries into 30 categories for the NQ and TQ,

[1] `https://github.com/subhadarship/kmeans_pytorch`

whereas for the WQ and TREC, we set the cluster number to 15. The number of keywords k is set to 20. Our experiments show the method is robust to cluster numbers (see Section 6.3). The key insight lies in grouping queries with similar semantic topics, not the exact cluster count.

When sampling queries from the same topic, we select two same-topic queries and passages to generate pseudo-passages, employing `GPT-3.5-turbo` for this task. In the projection of topic keywords, we obtain query vectors on all datasets using the official DPR-NQ checkpoint for an out-of-domain (OOD) investigation. We extract related keywords using the masked language model head of bert-base-uncased[2].

For DPR training, we adopt nearly identical experimental settings as DPR. However, due to computational constraints with only two A800 80G GPUs, we increase the batch size by a factor of 4 during training. For the NQ and Trivia, we train for 40 epochs, while for the TREC and WQ datasets, we train for 100 epochs. For the retriever inference, we follow DPR to use the English Wikipedia dump from Dec. 20, 2018 as the retrieval corpus. We perform all experiments five times using different random seeds and report the average accuracies.

5.1.3. Baselines

We compared our approach with multiple models.

- Contriever [49] is an unsupervised model, retrieves negative samples from context passages from previous batches.
- Spider [50] is a self-supervised learning approach used in information retrieval systems, it utilizes other passages in the same article for negative sampling.
- ART [51] retrieves a set of relevant evidence documents using a query and then reconstructs the query using the evidence documents.
- MSS-DPR[52] aims to refine the selection of passages over multiple stages, each stage potentially re-ranks or filters the output of the previous one based on increasing levels of relevance to the query.

[2] `https://huggingface.co/google-bert/bert-base-uncased`

Table 1. Main results on the four datasets report the retrieval accuracy at the top-20 and top-100 on the test sets. "Our impl." refers to our reproduction results of the original paper.

Model	NQ		TriviaQA		WQ		TREC	
	R@20	R@100	R@20	R@100	R@20	R@100	R@20	R@100
DPR [6]	78.4	85.4	79.4	85.0	73.2	81.4	79.8	89.1
Contriever [49]	67.8	82.1	74.2	83.2	-	-	-	-
Spider[50]	68.3	81.2	75.8	83.5	65.9	79.7	82.6	-
MSS-DPR [52]	81.4	88.1	81.9	86.6	76.9	**84.6**	-	-
DCSR [53]	78.9	86.5	79.7	85.2	-	-	-	-
ART [51]	80.2	88.4	**82.5**	**86.6**	74.4	82.7	-	-
Chain-of-Rewrite [54]	77.4	-	81.3	-	74.5	-	-	-
DPR(our impl.)	79.3	86.3	80.0	85.4	74.7	82.5	82.3	89.5
Query2doc [17](our impl.)	81.4	86.3	81.5	86.2	77.1	83.5	87.5	91.1
TDPR	**82.1**	**87.9**	82.3	86.2	**77.2**	83.8	**88.3**	**93.1**

- DCSR [53] implemented an in-passage negative sampling approach to promote the diverse generation of sentence representations within a single passage.
- Query2doc [17] generates pseudo-documents by sampling queries and passages randomly to train the dense document retriever.

5.2. Main Results

In the main experiment, we validate the retrieval accuracy of top-k (R@k) retrieved passages across four datasets, with the results presented in Table 1. TDPR outperforms DPR on all four datasets, achieving significant improvements at the top-20 accuracy ranging from 2.8% to 7.2%. Additionally, TDPR exhibits notable improvements over ART and MSS-DPR by 2.3% and 0.7%, respectively, at the top-20 accuracy on the NQ dataset. It maintains comparable performance to several baselines on the TriviaQA dataset and achieves around a 0.4% improvement over MSS-DPR at the top-20 accuracy on the WQ dataset. Furthermore, TDPR demonstrates enhanced retrieval accuracy compared to our implementation of Query2doc on all four datasets, with improvements ranging from 0.1% to 0.9% at the top-20 accuracy.

Table 2. Ablation Study of TDPR components on four datasets.

Model	NQ				TriviaQA				WQ			
	R@1	R@5	R@20	R@100	R@1	R@5	R@20	R@100	R@1	R@5	R@20	R@100
DPR(our impl.)	45.4	68.4	79.3	86.3	54.3	71.7	80.0	85.4	41.1	63.6	74.7	82.5
Query2doc [17](our impl.)	51.5	71.8	81.4	86.3	65.8	76.7	81.5	86.2	47.4	65.7	77.1	83.5
TDPR	**53.7**	**73.6**	82.1	**87.9**	67.0	77.1	**82.3**	86.2	47.8	**68.1**	77.2	**83.8**
-Keywords	53.4	72.2	**82.4**	87.7	**67.1**	**77.3**	82.3	**86.5**	**48.8**	67.7	**77.6**	83.7
-Pseudo-passage	45.6	67.9	79.7	86.1	53.8	70.8	79.4	84.9	41.4	64.9	76.0	83.3

6. Further Analysis

6.1. Ablation study

We perform ablation experiments on three datasets to assess the effectiveness of the individual modules outlined in our paper. As can be seen in Table 2, removing the topic-aware pseudo-passage ("-Pseudo-passage" in Table 2) results in varying degrees of accuracy decline. Meanwhile, it demonstrates that keywords after filtration lack substantial information, which can be validated via the keywords ablation ("-Keywords" in Table 2) results. This indicate that the performance improvement stems from the complementary effects of pseudo-passages and keywords. The pseudo-passages generated through topic-aware ICL provide contextual information (e.g., entity relationships in Tokyo Imperial Palace case in Fig. 1), while the keywords expand lexical coverage through MLM-based projection (e.g., "california" related terms in Table 6). This dual approach addresses both semantic gaps (through structured passages) and vocabulary mismatches (through domain-specific keywords), confirming that removing either component can leads to performance degradation, indicating their synergistic effects. It is worth mentioning that using query representations generated by the official DPR-NQ for vocabulary projection may bring noise on out-of-domain (OOD) three datasets. On the other hand, the query representation of the OOD retriever can still provide potential topic keywords to a certain extent, highlighting the possibility of not having to pre-train a retriever for each dataset. Additionally, the varying improvements across datasets Table 1 correlate with their inherent characteristics. TREC's questions benefit most from topic-aware expansion (+6% R@20) due to their implicit contextual needs, while Triv-

Table 3. Impact of different sampling strategies on retrieval accuracy.

Model	NQ				TQ				WQ			
	R@1	R@5	R@20	R@100	R@1	R@5	R@20	R@100	R@1	R@5	R@20	R@100
DPR(our impl.)	45.4	68.4	79.3	86.3	54.3	71.7	80.0	85.4	41.1	63.6	74.7	82.5
TDPR	**53.7**	**73.6**	**82.1**	**87.9**	**67.0**	**77.1**	**82.3**	**86.2**	**47.8**	**68.1**	**77.2**	83.8
TDPR(zero shot)	52.1	72.5	81.9	87.2	65.7	76.4	81.7	86.0	46.4	66.2	77.3	83.6
TDPR(random)	51.4	72.2	81.2	87.5	65.6	76.4	81.8	86.4	45.0	66.1	76.8	83.7
TDPR(topic random)	51.4	73.1	81.3	87.1	65.5	76.3	82.0	86.3	43.2	65.2	75.4	**83.9**

iaQA's questions show smaller gains (+2.3% R@20) as original queries are typically factual, and the answers often appear directly in the relevant documents. Therefore, if the original queries are already detailed enough, additional keywords may not be necessary to boost the retrieval accuracy. Moreover, it is interesting that when pseudo-passages and keywords are combined, noise from irrelevant keywords seems to be diluted. We hypothesize that frequently appearing relevant words in pseudo-passages compensate for the error, which can be confirmed by referring to [17] enhancing sparse retrieval such as BM25 through repetition.

6.2. Different Sampling Strategy

In Table 3, we examine the impact of different sampling strategies on ICL. We use four distinct sampling strategies: the proposed topic-aware sampling, random sampling from the entire dataset, random sampling from different topic clusters, and a zero-shot setting. The results show that our model performs best in the topic-aware setting. Although zero-shot, random sampling, and random sampling from topic clusters also significantly outperform the original DPR, their performance in top-1 accuracy decreases successively. This aligns with our intuition that the topic-aware approach helps the model understand how to generate pseudo-passage from queries within the same topic, aiding retrieval. Random sampling, since it involves selecting queries and passages randomly from the entire dataset, may choose queries from different topics, impacting the generation of pseudo-passage. Lastly, random sampling from topic clusters, as it likely selects queries most dissimilar to the original ones, reflects the poorest retrieval performance among these four settings.

6.3. Topic Coverage

In this section, we further investigate the impact of topic coverage on retrieval performance through clustering and keywords. In Table 4, we begin by comparing the performance across various cluster counts ($n = 5; 15; 25; 30; 35$). We can see that: (1) The number of clusters under which the model achieves optimal performance varies on different datasets. This is because the datasets are collected from different sources and have different distributions; (2) The model performs best when $n = 15$ on TREC and WQ and $n = 30$ on TQ and NQ, but the margin leading $n = \{25, 30\}$ is not significant. When $n = 5$, the performance drops by a large margin. We hypothesize that fewer topic categories may introduce irrelevant demonstrations and make it challenging to play a positive role in pseudo-passage generation; (3) Excessive topic division may be unnecessary or even have a negative impact (e.g., 64.6% w/ $n = 15$ vs. 63.8% w/ $n = 35$ on TREC). We hypothesize that the datasets do not have such a number of individual clusters. As a result, the clustering algorithm is hard to converge. The performance remains stable within a range of cluster numbers ($n = 20$ to 35 for NQ and TQ, $n = 10$ to 20 for WQ and TREC), which indicates the effectiveness is driven by topic clustering itself, not specific n-values. Although cluster numbers require empirical experience, we emphasize the effectiveness of clustering and that n can be determined by automated methods. In practice, for datasets with broad topics (e.g., NQ), use bigger n can ensure sufficient examples per cluster while smaller n values can balance topic granularity and demonstration quality on datasets with narrower topics such as WQ. In addition, Fig. 3 depicts the distribution visualization of NQ and TQ training problems with different numbers of clusters. We can observe that both insufficient and excessive topic segmentation can lead to imbalanced demonstration sampling. For example, as shown in Fig. 3(a) and Fig. 3(d), a large number of queries are scattered far away from their own centroids and even intertwined with samples from other clusters, introducing noising demonstrations during sampling. Meanwhile, Fig. 3(c) and Fig. 3(f) illustrate examples of over-clustering leading to insufficient representativeness in some clusters. Therefore, the uniform distribution of topics significantly impacts the sampling of examples and the quality of LLMs generation.

Then, we explore the role of keywords. Since the effect of word frequency on

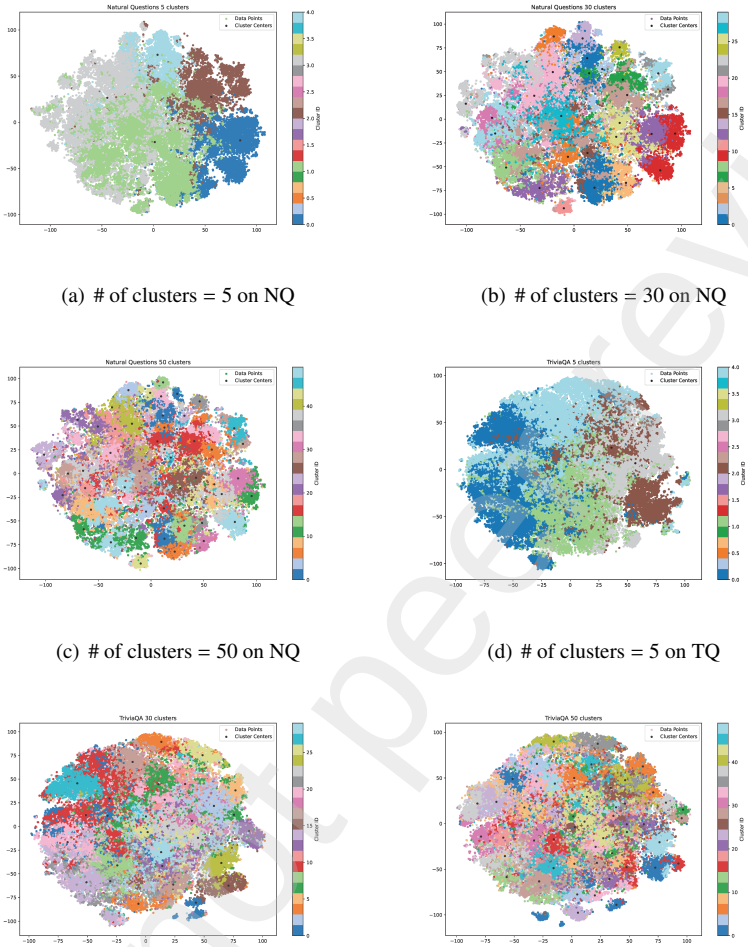
Table 4. Retrieval accuracy (%) on the NQ, TQ, WQ and TREC dataset with the different setting of clustering and keywords. We use the query representations of DPR-NQ for keyword projection.

Method	NQ				TQ				WQ				TREC			
	R@1	R@5	R@20	R@100	R@1	R@5	R@20	R@100	R@1	R@5	R@20	R@100	R@1	R@5	R@20	R@100
TDPR	**53.9**	73.4	81.9	**88.0**	**67.3**	77.2	**82.5**	**86.5**	47.0	67.0	77.2	**84.1**	**66.7**	80.9	**88.6**	**93.8**
clustering																
# clusters = 5	52.3	72.6	81.3	87.2	65.6	76.3	81.1	85.6	46.7	67.3	76.5	82.9	63.1	81.6	87.2	92.4
# clusters = 15	53.0	73.0	81.9	87.1	66.3	76.4	81.8	85.7	-	-	-	-	-	-	-	-
# clusters = 25	53.4	73.0	81.9	88.0	66.7	**77.3**	81.9	86.1	47.8	67.7	**77.4**	83.5	64.3	82.3	88.0	93.2
# clusters = 30	-	-	-	-	-	-	-	-	47.4	67.8	76.8	83.1	64.3	81.8	88.1	92.9
# clusters = 35	53.8	73.3	**82.4**	87.7	66.5	76.7	82.3	85.8	47.1	67.4	76.8	83.0	63.8	81.3	87.7	92.6
keywords																
Remove overlapping keywords	53.7	**73.6**	82.1	87.9	67.0	77.1	82.3	86.2	**47.8**	**68.1**	77.2	83.8	64.6	**82.4**	88.3	93.1
Only remove tokens existing in q	53.9	73.3	82.0	87.9	67.3	77.1	82.4	86.5	47.2	67.3	76.8	84.0	66.3	81.3	88.5	93.6
Only remove tokens existing in $d+$	53.2	73.5	82.1	87.8	67.1	77.2	82.0	86.3	47.7	68.1	77.3	83.6	64.7	82.0	88.5	93.2

dense search is not significant compared to sparse search, we find that keeping all projected tokens has no consistent effect across different datasets. In particular, overlapping keywords boost the top-1 retrieval performance on NQ, TQ, and TREC but not other metrics, demonstrating a similar conclusion that there is a trade-off between topic relevance (through repetition) and coverage (through diversity) [41]. For WQ, the query encoder of the NQ-based retriever reflects poor OOD capabilities. Hence, irrelevant overlapping keywords have a negative impact on the ranking of gold passages. Furthermore, the impact of only removing keywords in q is insignificant, while deleting those in pseudo-passages has a noticeable change on NQ, TQ and TREC. A possible reason is that the pseudo-passage generated by LLM contains more potential topics than the shorter query. Hallucinations from pseudo-passages and errors from vocabulary projection may be another reasonable explanation.

6.4. LLM Scale

We evaluate the impact of LLM of different sizes on retrieval performance on two datasets. Our default setting used `GPT-3.5-turbo` for pseudo-passage generation. In addition, we tested the accuracy of the retrieval model on LLM of different sizes, specifically on the `LLaMA-7b` and `LLaMA-13b` settings, also applying the topic-aware setting for both. As shown in Fig. 4, although models of different sizes exhibit varying retrieval performances, the pseudo-passages generated by all contribute to an improvement in accuracy. From this, we can infer that although the generated pseudo-passages



(a) # of clusters = 5 on NQ

(b) # of clusters = 30 on NQ

(c) # of clusters = 50 on NQ

(d) # of clusters = 5 on TQ

(e) # of clusters = 30 on TQ

(f) # of clusters = 50 on TQ

Fig. 3. Visualization of the topic distribution for training queries on NQ and TQ.

may not necessarily contain the correct answer, the overall content of these passages helps the model in retrieving documents that do contain the correct answer.

6.5. Time Cost

We also analyze the time cost (ms/q) during the online inference phase on two datasets. Following Query2doc, we use the API of `GPT-3.5-turbo` to retrieve top-20

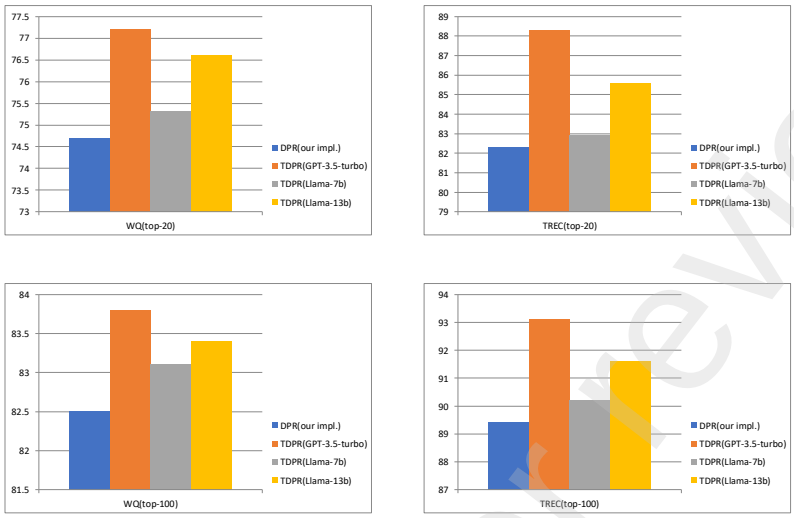


Fig. 4. Effect of using pseudo-passages generated by LLMs of different scales on retrieval accuracy.

passages with a single thread. As shown in Table 5, the time delay caused by generating pseudo-passages accounts for 87%-89% of the inference time compared with the vanilla DPR. Instead, the time required for matching test queries on unsupervised clustering is trivial, similar to vocabulary projection. Different from the sparse retrieval reported in Query2doc, dense retrieval is insensitive to query length, so the time required for index search is almost the same as that required for DPR. Although the time consumption of pseudo-passage generation mainly depends on server load and network latency, it is necessary to consider how to balance efficiency and effectiveness for real-world scenarios.

6.6. Case Study

Observing the two examples in Table 6, we conclude that the pseudo-passages generated in both examples effectively contain key information that may exist in the ground-truth passages. Additionally, the keywords contain information relevant to the queries, such as “California” and “Mexico” in the first example and “Europe” and “Brazil” in the second example. While some keywords may not directly appear in the

Table 5. Time cost (ms/q) of online operations of TDPR. We use the API of `GPT-3.5-turbo` to retrieve top-20 passages with a single thread.

Online Operations	NQ		WQ	
	DPR	TDPR	DPR	TDPR
Avg. query Clustering	-	0.4	-	0.4
Avg. Demonstration Selection	-	-	-	-
Avg. query Projection	-	1.3	-	1.5
Avg. Per Pseudo-passage Generation	-	488	-	405
Avg. query Encoding	0.7	1.1	0.6	0.8
Avg. Retrieval	52	53	55	55

ground-truth passages, they are still directly relevant to the queries. This demonstrates that using BERT to parse DPR's query vectors can explicitly capture implicit knowledge in the DPR model. Furthermore, although we removed words already present in the generated pseudo-passages and original queries when using keywords, we can still obtain highly relevant keywords to the ground-truth passages. The keywords and generated pseudo-passages complement each other well.

7. Conclusion

This study introduces a simple yet effective method for augmenting open-domain dense retrieval. By concatenating topic-aware ICL and topic keywords projection with the original query, we aim to augment the information of the original query. Experiment results show that our proposed method achieves competitive performance on four ODQA datasets compared to recent SOTA studies. Additionally, we conduct extensive experiments to demonstrate the effectiveness of our proposed method. We also perform experiments using different sampling methods to verify the impact of large language models of different sizes on the sampling. Furthermore, we analyze the time cost of this study. Although our study improves retrieval performance, there is an addi-

Table 6. Example from the WQ dataset. The red part indicates keywords that are relevant to the original query but do not appear in either the pseudo-passages or any of the ground-truth passages.

query	what to see near sedona arizona?
keywords	ca california west something town se nothing la mexico go good one place north 1 city day back said south
pseudo-passages	Sedona, Arizona is known for its beautiful red rock formations, stunning hiking trails, and art galleries. Visitors can explore the iconic Cathedral Rock, hike the scenic Bell Rock Trail, and visit the art and craft shops in the Tlaquepaque Arts and Crafts Village. There are also vortex sites for spiritual healing and reflection, as well as jeep tours for a unique way to experience the natural beauty of the area.
ground-truth passage	unincorporated Village of Oak Creek, to the south and well outside the Sedona city limits, is a significant part of the Sedona community. Sedona became one of the Arizona municipalities to approve of civil unions for same-sex partners. The Sedona area hosts numerous events annually, including: Sedona hosts several notable arts organizations in Northern Arizona: A specialized New Age tourist industry operates in Sedona, where Jose Arguelles organized the "Harmonic Convergence" in 1987. Some New Age proponents purport that "spiritual vortices" are concentrated in the Sedona area at Bell Rock, Airport Mesa, Cathedral Rock, and Boynton Canyon.
query	what countries do people speak portuguese?
keywords	language portugal languages brazil nationality spain country islands spanish french italy english one different island europe others england east
pseudo-passages	Portuguese is the official language of Portugal, Brazil, Mozambique, Angola, Guinea-Bissau, East Timor, Equatorial Guinea, and Sao Tome and Principe. It is also spoken in Macau, a territory of China, and in Goa, Daman, and Diu in India. Additionally, there are Portuguese-speaking communities in various other countries around the world.
ground-truth passage	Portuguese overseas territory to be decolonized, and returned to China in 1999. It still retains traces of the Portuguese culture and Portuguese is one of the official languages of the territory. Despite that, the majority of the population in Macau do not speak and understand Portuguese. Cantonese language is the main language. In 2006, during the II Ministerial meeting between China and Portuguese Speaking Countries, the CPLP Executive Secretary and Deputy ambassador Tadeu Soares invited the Chief Executive of the Government of the Macau Special Administrative Region, Edmund Ho Hau Wa, to request the Associate Observer status for Macau.

tional time cost during inference due to the need for large language models to generate pseudo-passages. In future work, we can consider reducing latency while expanding the practical information for the query.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China (Grant No. 62302354).

References

[1] O. Ram, L. Bezalel, A. Zicher, Y. Belinkov, J. Berant, A. Globerson, What are you token about? dense retrieval as distributions over the vocabulary, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, 2023, pp. 2481–2498. URL: https://doi.org/10.18653/v1/2023.acl-long.140. doi:10.18653/V1/2023.ACL-LONG.140.

[2] P. S. H. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

[3] Z. Jiang, F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, G. Neubig, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, 2023, pp. 7969–7992. URL: https://doi.org/10.18653/v1/2023.emnlp-main.495. doi:10.18653/V1/2023.EMNLP-MAIN.495.

[4] S. Siriwardhana, R. Weerasekera, T. Kaluarachchi, E. Wen, R. Rana, S. Nanayakkara, Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering, Trans. Assoc. Comput. Linguistics 11 (2023) 1–17. URL: https://transacl.org/ojs/index.php/tacl/article/view/4029.

[5] E. Melz, Enhancing LLM intelligence with ARM-RAG: auxiliary rationale memory for retrieval augmented generation, CoRR abs/2311.04177 (2023). URL: https://doi.org/10.48550/arXiv.2311.04177. doi:10.48550/ARXIV.2311.04177. arXiv:2311.04177.

[6] V. Karpukhin, B. Oguz, S. Min, P. S. H. Lewis, L. Wu, S. Edunov, D. Chen, W. Yih, Dense passage retrieval for open-domain question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, 2020, pp. 6769–6781. URL: https://doi.org/10.18653/v1/2020.emnlp-main.550. doi:10.18653/V1/2020.EMNLP-MAIN.550.

[7] Y. Qu, Y. Ding, J. Liu, K. Liu, R. Ren, W. X. Zhao, D. Dong, H. Wu, H. Wang, Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, 2021, pp. 5835–5847. URL: https://doi.org/10.18653/v1/2021.naacl-main.466. doi:10.18653/V1/2021.NAACL-MAIN.466.

[8] L. Gao, J. Callan, Condenser: a pre-training architecture for dense retrieval, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, 2021, pp. 981–993. URL: https://doi.org/10.18653/v1/2021.emnlp-main.75. doi:10.18653/V1/2021.EMNLP-MAIN.75.

[9] L. Gao, J. Callan, Unsupervised corpus aware language model pre-training for dense passage retrieval, in: Proceedings of the 60th Annual Meeting

of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, 2022, pp. 2843–2853. URL: https://doi.org/10.18653/v1/2022.acl-long.203. doi:10.18653/V1/2022.ACL-LONG.203.

[10] L. Xiong, C. Xiong, Y. Li, K. Tang, J. Liu, P. N. Bennett, J. Ahmed, A. Overwijk, Approximate nearest neighbor negative contrastive learning for dense text retrieval, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, 2021. URL: https://openreview.net/forum?id=zeFrfgyZln.

[11] K. Wang, N. Thakur, N. Reimers, I. Gurevych, GPL: generative pseudo labeling for unsupervised domain adaptation of dense retrieval, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022, 2022, pp. 2345–2360. URL: https://doi.org/10.18653/v1/2022.naacl-main.168. doi:10.18653/V1/2022.NAACL-MAIN.168.

[12] V. Lavrenko, W. B. Croft, Relevance-based language models, in: SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA, 2001, pp. 120–127. URL: https://doi.org/10.1145/383952.383972. doi:10.1145/383952.383972.

[13] S. Zhang, W. Fan, B. He, CBIA VT at TREC 2015 clinical decision support track - exploring relevance feedback and query expansion in biomedical information retrieval, in: Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015, volume 500-319 of *NIST Special Publication*, 2015. URL: http://trec.nist.gov/pubs/trec24/papers/CBIA_VT-CL.pdf.

[14] J. Wang, M. Pan, T. He, X. Huang, X. Wang, X. Tu, A pseudo-relevance feedback framework combining relevance matching and semantic matching for information

retrieval, Inf. Process. Manag. 57 (2020) 102342. URL: https://doi.org/10.1016/j.ipm.2020.102342. doi:10.1016/J.IPM.2020.102342.

[15] OpenAI, GPT-4 technical report, CoRR abs/2303.08774 (2023). URL: https://doi.org/10.48550/arXiv.2303.08774. doi:10.48550/ARXIV.2303.08774. arXiv:2303.08774.

[16] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, CoRR abs/2302.13971 (2023). URL: https://doi.org/10.48550/arXiv.2302.13971. doi:10.48550/ARXIV.2302.13971. arXiv:2302.13971.

[17] L. Wang, N. Yang, F. Wei, Query2doc: Query expansion with large language models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, 2023, pp. 9414–9423. URL: https://doi.org/10.18653/v1/2023.emnlp-main.585. doi:10.18653/V1/2023.EMNLP-MAIN.585.

[18] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, W. Chen, What makes good in-context examples for gpt-3?, in: Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, DeeLIO@ACL 2022, Dublin, Ireland and Online, May 27, 2022, 2022, pp. 100–114. URL: https://doi.org/10.18653/v1/2022.deelio-1.10. doi:10.18653/V1/2022.DEELIO-1.10.

[19] K. Peng, L. Ding, Y. Yuan, X. Liu, M. Zhang, Y. Ouyang, D. Tao, Revisiting demonstration selection strategies in in-context learning, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, 2024, pp. 9090–9101. URL: https://aclanthology.org/2024.acl-long.492.

[20] V. M. S., M. Van, X. Wu, In-context learning demonstration selection via influence analysis, CoRR abs/2402.11750 (2024). URL: https://doi.

org/10.48550/arXiv.2402.11750. doi:10.48550/ARXIV.2402.11750. arXiv:2402.11750.

[21] E. M. Voorhees, D. M. Tice, The TREC-8 question answering track, in: Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000, 31 May - June 2, 2000, Athens, Greece, 2000. URL: http://www.lrec-conf.org/proceedings/lrec2000/html/summary/26.htm.

[22] Wikipedia, Wikipedia, 2004.

[23] Y. Zhu, R. Kiros, R. S. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, S. Fidler, Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, in: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, 2015, pp. 19–27. URL: https://doi.org/10.1109/ICCV.2015.11. doi:10.1109/ICCV.2015.11.

[24] J. Devlin, M. Chang, K. Lee, K. Toutanova, 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/V1/N19-1423.

[25] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020) 140:1–140:67. URL: https://jmlr.org/papers/v21/20-074.html.

[26] S. MacAvaney, F. M. Nardini, R. Perego, N. Tonellotto, N. Goharian, O. Frieder, Expansion via prediction of importance with contextualization, in: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020,

2020, pp. 1573–1576. URL: https://doi.org/10.1145/3397271.3401262. doi:10.1145/3397271.3401262.

[27] L. Adolphs, M. C. Huebscher, C. Buck, S. Girgin, O. Bachem, M. Ciaramita, T. Hofmann, Decoding a neural retriever's latent space for query suggestion, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, 2022, pp. 8786–8804. URL: https://doi.org/10.18653/v1/2022.emnlp-main.601. doi:10.18653/V1/2022.EMNLP-MAIN.601.

[28] Q. Xiao, S. Li, L. Chen, Topic-dpr: Topic-based prompts for dense passage retrieval, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023, 2023, pp. 7216–7225. URL: https://doi.org/10.18653/v1/2023.findings-emnlp.480. doi:10.18653/V1/2023.FINDINGS-EMNLP.480.

[29] W. Yu, D. Iter, S. Wang, Y. Xu, M. Ju, S. Sanyal, C. Zhu, M. Zeng, M. Jiang, Generate rather than retrieve: Large language models are strong context generators, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, 2023. URL: https://openreview.net/pdf?id=fB0hRu9GZUS.

[30] G. A. Miller, WORDNET: a lexical database for english, in: Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, USA, February 23-26, 1992, 1992. URL: https://aclanthology.org/H92-1116/.

[31] A. Chakraborty, D. Ganguly, O. Conlan, Retrievability based document selection for relevance feedback with automatically generated query variants, in: CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020, 2020, pp. 125–134. URL: https://doi.org/10.1145/3340531.3412032. doi:10.1145/3340531.3412032.

[32] Z. Zheng, K. Hui, B. He, X. Han, L. Sun, A. Yates, Contextualized query expansion via unsupervised chunk selection for text retrieval, Inf. Process. Manag.

58 (2021) 102672. URL: https://doi.org/10.1016/j.ipm.2021.102672. doi:10.1016/J.IPM.2021.102672.

[33] Y. Lv, C. Zhai, A comparative study of methods for estimating query language models with pseudo feedback, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009, 2009, pp. 1895–1898. URL: https://doi.org/10.1145/1645953.1646259. doi:10.1145/1645953.1646259.

[34] X. Wang, Z. Wang, X. Gao, F. Zhang, Y. Wu, Z. Xu, T. Shi, Z. Wang, S. Li, Q. Qian, R. Yin, C. Lv, X. Zheng, X. Huang, Searching for best practices in retrieval-augmented generation, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024, 2024, pp. 17716–17736. URL: https://aclanthology.org/2024.emnlp-main.981.

[35] S. Dai, C. Xu, S. Xu, L. Pang, Z. Dong, J. Xu, Bias and unfairness in information retrieval systems: New challenges in the LLM era, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024, 2024, pp. 6437–6447. URL: https://doi.org/10.1145/3637528.3671458. doi:10.1145/3637528.3671458.

[36] S. Dai, Y. Zhou, L. Pang, W. Liu, X. Hu, Y. Liu, X. Zhang, G. Wang, J. Xu, Neural retrievers are biased towards llm-generated content, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024, 2024, pp. 526–537. URL: https://doi.org/10.1145/3637528.3671882. doi:10.1145/3637528.3671882.

[37] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T. Chua, Q. Li, A survey on RAG meeting llms: Towards retrieval-augmented large language models, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024,

2024, pp. 6491–6501. URL: https://doi.org/10.1145/3637528.3671470. doi:10.1145/3637528.3671470.

[38] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[39] L. Gao, X. Ma, J. Lin, J. Callan, Precise zero-shot dense retrieval without relevance labels, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, 2023, pp. 1762–1777. URL: https://doi.org/10.18653/v1/2023.acl-long.99. doi:10.18653/V1/2023.ACL-LONG.99.

[40] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL: http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

[41] H. Su, J. Kasai, C. H. Wu, W. Shi, T. Wang, J. Xin, R. Zhang, M. Ostendorf, L. Zettlemoyer, N. A. Smith, T. Yu, Selective annotation makes language models better few-shot learners, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, 2023. URL: https://openreview.net/pdf?id=qY1hlv7gwg.

[42] O. Rubin, J. Herzig, J. Berant, Learning to retrieve prompts for in-context learning, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022, 2022, pp. 2655–2671. URL: https://doi.org/10.18653/v1/2022.naacl-main.191. doi:10.18653/V1/2022.NAACL-MAIN.191.

[43] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with gpus, IEEE Trans. Big Data 7 (2021) 535–547. URL: https://doi.org/10.1109/TBDATA.2019.2921572. doi:10.1109/TBDATA.2019.2921572.

[44] P. Lewis, Y. Wu, L. Liu, P. Minervini, H. Küttler, A. Piktus, P. Stenetorp, S. Riedel, PAQ: 65 million probably-asked questions and what you can do with them, Trans. Assoc. Comput. Linguistics 9 (2021) 1098–1115. URL: https://doi.org/10.1162/tacl_a_00415. doi:10.1162/TACL_A_00415.

[45] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. P. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M. Chang, A. M. Dai, J. Uszkoreit, Q. Le, S. Petrov, Natural questions: a benchmark for question answering research, Trans. Assoc. Comput. Linguistics 7 (2019) 452–466. URL: https://doi.org/10.1162/tacl_a_00276. doi:10.1162/TACL_A_00276.

[46] M. Joshi, E. Choi, D. S. Weld, L. Zettlemoyer, Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, 2017, pp. 1601–1611. URL: https://doi.org/10.18653/v1/P17-1147. doi:10.18653/V1/P17-1147.

[47] J. Berant, A. Chou, R. Frostig, P. Liang, Semantic parsing on freebase from question-answer pairs, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT,

a Special Interest Group of the ACL, 2013, pp. 1533–1544. URL: https://aclanthology.org/D13-1160/.

[48] P. Baudis, J. Sedivý, Modeling of the question answering task in the yodaqa system, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 6th International Conference of the CLEF Association, CLEF 2015, Toulouse, France, September 8-11, 2015, Proceedings, volume 9283 of *Lecture Notes in Computer Science*, 2015, pp. 222–228. URL: https://doi.org/10.1007/978-3-319-24027-5_20. doi:10.1007/978-3-319-24027-5_20.

[49] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, E. Grave, Unsupervised dense information retrieval with contrastive learning, Trans. Mach. Learn. Res. 2022 (2022). URL: https://openreview.net/forum?id=jKN1pXi7b0.

[50] O. Ram, G. Shachaf, O. Levy, J. Berant, A. Globerson, Learning to retrieve passages without supervision, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022, 2022, pp. 2687–2700. URL: https://doi.org/10.18653/v1/2022.naacl-main.193. doi:10.18653/V1/2022.NAACL-MAIN.193.

[51] D. S. Sachan, M. Lewis, D. Yogatama, L. Zettlemoyer, J. Pineau, M. Zaheer, Questions are all you need to train a dense passage retriever, Trans. Assoc. Comput. Linguistics 11 (2023) 600–616. URL: https://doi.org/10.1162/tacl_a_00564. doi:10.1162/TACL_A_00564.

[52] D. S. Sachan, M. Lewis, M. Joshi, A. Aghajanyan, W. Yih, J. Pineau, L. Zettlemoyer, Improving passage retrieval with zero-shot question generation, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, 2022, pp. 3781–3797. URL: https://doi.org/10.18653/v1/2022.emnlp-main.249. doi:10.18653/V1/2022.EMNLP-MAIN.249.

[53] W. Hong, Z. Zhang, J. Wang, H. Zhao, Sentence-aware contrastive learning for open-domain passage retrieval, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, 2022, pp. 1062–1074. URL: https://doi.org/10.18653/v1/2022.acl-long.76. doi:10.18653/V1/2022.ACL-LONG.76.

[54] C. Xin, Y. Lu, H. Lin, S. Zhou, H. Zhu, W. Wang, Z. Liu, X. Han, L. Sun, Chain-of-rewrite: Aligning question and documents for open-domain question answering, in: Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024, 2024, pp. 1884–1896. URL: https://aclanthology.org/2024.findings-emnlp.104.