Unsupervised Text Representation Learning via Instruction-Tuning for Zero-Shot Dense Retrieval

Qiuhai Zeng[*][†]
The Pennsylvania State University
qjz5084@psu.edu

Zimeng Qiu[*]
Amazon AGI
zimengqi@amazon.com

Dae Yon Hwang[*]
Amazon AGI
dyhwang@amazon.com

Xin He
Amazon AGI
xih@amazon.com

William M. Campbell
Amazon AGI
cmpw@amazon.com

Abstract

Dense retrieval systems are commonly used for information retrieval (IR). They rely on learning text representations through an encoder and usually require supervised modeling via labelled data which can be costly to obtain or simply unavailable. In this study, we introduce a novel unsupervised text representation learning technique via instruction-tuning the pre-trained encoder-decoder large language models (LLM) under the dual-encoder retrieval framework. We demonstrate the corpus representation can be augmented by the representations of relevant synthetic queries generated by the instruct-tuned LLM founded on the Rao-Blackwell theorem. Furthermore, we effectively align the query and corpus text representation with self-instructed-tuning. Specifically, we first prompt an open-box pre-trained LLM to follow defined instructions (i.e. question generation and keyword summarization) to generate synthetic queries. Next, we fine-tune the pre-trained LLM with defined instructions and the generated queries that passed quality check. Finally, we generate synthetic queries with the instruction-tuned LLM for each corpora and represent each corpora by weighted averaging the synthetic queries and original corpora embeddings. We evaluate our proposed method under low-resource settings on three English and one German retrieval datasets measuring NDCG@10, MRR@100, Recall@100. We significantly improve the average zero-shot retrieval performance on all metrics, increasing open-box FLAN-T5 model variations by $[3.34\%, 3.50\%]$ in absolute and exceeding three competitive dense retrievers (i.e. mDPR, T-Systems, mBART-Large), with model of size at least 38% smaller, by 1.96%, 4.62%, 9.52% absolute on NDCG@10.

CCS Concepts

• **Information systems → Novelty in information retrieval**.

Keywords

Instruction-tuning, Zero-shot, Unsupervised Data Augmentation, Dense Retrieval

ACM Reference Format:
Qiuhai Zeng, Zimeng Qiu, Dae Yon Hwang, Xin He, and William M. Campbell. 2018. Unsupervised Text Representation Learning via Instruction-Tuning for Zero-Shot Dense Retrieval. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Dense retrieval systems commonly employ dual-encoder retrieval models which use two separate encoders, either symmetric or asymmetric, to represent the query and corpus [8, 9, 13, 33]. The corpora are indexed with representation and will be retrieved in response to each query based on the relevance scores. The scores are usually calculated based on embedding similarity, such as dot product or cosine similarity. Although dense retrieval systems have developed rapidly, the model performance largely depends supervised text representation learning and relevancy capturing between the query and corpus [36]. Yet, it remains to be a major challenge to properly retrieve when lacking labeled modeling data. Existing work [21, 22] leveraged pre-trained large encoders (specifically T5 models, Raffel et al. [26]) to alleviate the data thirst. However, their proposals still required annotated datasets either by web mining or manual annotation for fine-tuning in order to improve the generalization ability of dual-encoder retrieval models, for example, dealing with out-of-domain data. An alternative solution is to train a dense retrieval on synthetic query-corpus relevance pairs. [18] trains a question generation system on general domain data and applies it to the targeted domain to construct synthetic question-passage data. To save the effort of training a task-specific generation model on general data, like Natural Questions [15] or MSMARCO [20], Promptagator [5] proposes to use pre-trained LLMs, like FLAN [32], as a few-shot query generator to build the data for training the dual-encoder. However, the synthetic queries are not directly leveraged at inference, potentially causing gaps between training and inference of dense retrievers [2]. Earlier work, e.g., doc2query [24], concatenates the generated queries with the corresponding corpus, aiming to enrich the corpus representation with questions that the corpus can potentially answer. An improved version, docTTTTTquery [23] leverages pre-trained T5 models as the expansion model, leading to more relevant synthetic queries and better retrieval performance.

[*] Those authors contributed equally to this research.
[†] Work done while intern at Amazon.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

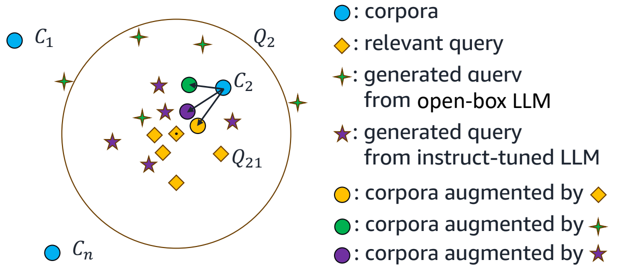
C_1
Q_2
C_2
Q_{21}
C_n
●: corpora
◆: relevant query
✦: generated query from open-box LLM
★: generated query from instruct-tuned LLM
●: corpora augmented by ◆
●: corpora augmented by ✦
●: corpora augmented by ★

Figure 1: Illustration of the corpus representation augmented by the embedding of relevant queries, synthetic queries generated by open-box LLM and instruct-tuned LLM.

Different from the previous work, we demonstrate directly on the embedding level instead of the text level, that the synthetically generated queries' embeddings can effectively augment the corpus representation (Figure 1). Here, we propose an unsupervised representation learning approach through self-instructed-tuning leveraging the embedding generation and sequence generation capability of an encoder-decoder LLM. This approach consists of two steps, i.e. self-instructed-learning and Rao-Blackwellization. In the first step, we design two instruction tasks, namely question generation and keyword summarization, to generate synthetic questions and keywords for each given corpus via prompting a pre-trained LLM. Next, we apply filters to gate the synthetic data quality and instruction-tune the pre-trained LLM (and its variant versions) on the filtered output (Step one in Figure 2). In the second step, we use the instruct-tuned LLM to generate better synthetic questions and keywords following the same instruction prompts as in training. We then obtain the embeddings of the newly generated synthetic questions and keywords and that of the corpus from the instruct-tuned LLM encoder, and take the weighted average as our augmented corpus representation (Step two in Figure 2).

We consider the corpus representation learning task as a problem of query embedding expectation estimation. Based on the Rao-Blackwell theorem, the crude estimator, corpus embedding, can be improved by taking the conditional expectation given the sufficient statistics, i.e. sample mean of the embedding of their (synthetic) relevant queries and keywords. Thus, we expect combining the raw corpus embedding and synthetic query embedding to achieve better corpus representation. Besides, by aligning instruction-tuning and synthetic query generation, the retrieval model is directly optimized on corpus representation during training. To assess the effectiveness of our proposed method, we compare retrieval method of corpus only embedding with our augmented corpus representation, models with and without instruction-tuning and evaluate against multiple competitive dense retrievers (i.e., mBART [29], mDPR [34, 35], T-Systems [28]). Our main contributions are as follows:

- We propose a novel unsupervised text representation learning approach for dual-encoder retrieval model by instruction-tuning a pre-trained encoder-decoder using unlabelled corpus.
- We demonstrate our approach of using conditional expectation of the relevant (synthetic) query/keywords embedding the representation of the corpus can be augmented effectively, founded on the Rao-Blackwell theorem.
- We verify the effectiveness of the proposed methods on three English and one German IR datasets measured by NDCG@10, MRR@100, Recall@100. We significantly improve the zero-shot average retrieval performance on all metrics with our unsupervised approach and exceed three competitive supervised dense retrievers, with model of size at least 38% smaller, by 1.96%, 4.62%, 9.52% absolute on NDCG@10 (Table 4).

2 Related Work

2.1 Instruction-tuning

Tuning pre-trained LLMs with *(natural language instruction, response)* pairs to enhance models' ability to follow instructions and understand user intention. It is a rising paradigm in NLP to strengthen model's generalizability on unseen tasks. FLAN [32] significantly improves a 137B LLM's zero-shot performance via instruction learning on various NLP datasets with multiple instruction templates. InstructDial [10] also shows significant zero-shot performance boost in unseen dialogues when applying instruction-tuning to dialogue domain. InstructGPT [25] enhances GPT-3's performance by fine-tuning it on instructions and human feedback collected from OpenAI API. Self-Instruct [31] fine-tunes the open-box GPT-3 on its own generated instructions and instances which achieved on par performance of InstructGPT.

2.2 Dense Retrieval Text Representation

The foundational component of dense retrieval is the text representation. Under dual-encoder framework, it has been a long standing practice such as Sentence-BERT [27], ColBERT [14] to represent query and corpus with encoder only models, e.g., BERT [7] and RoBERTa [16]. Recently Sentence-T5 [21] demonstrate that encoder-decoder pretrained LLM like T5 can achieve superior performance. Further, representing corpus with single representation may not well model the fine-grained semantic interaction between the queries and corpus. Poly-encoder [11] and ME-BERT [17] learn multiple representations to better capture the corpus semantics and show significant improvement. Doc2query [24] and docTTTTTquery [23] append generated synthetic queries to the corpus and thus enrich the semantic information.

3 Method

We propose an unsupervised text representation learning approach through self-instructed-tuning a pre-trained encoder-decoder LLM. The first step is to generate instruction following responses from an LLM and instruction-tune the LLM itself with filtered quality *(natural language instruction, response)* pairs. The second step is to compute the augmented corpus embedding weighing in synthetic queries' (e.g. questions, keywords) embeddings. Figure 2 presents the overall flow of our approach.

3.1 Problem Scenario

Denote corpora as C_1,C_2,...,C_n, and their relevant queries as Q_{11}, Q_{12},...,Q_{21},..., where queries Q_{i1},Q_{i2},... are relevant to the same corpora C_i. For example, Q_{11} can be `Harry Potter 1` and Q_{12}

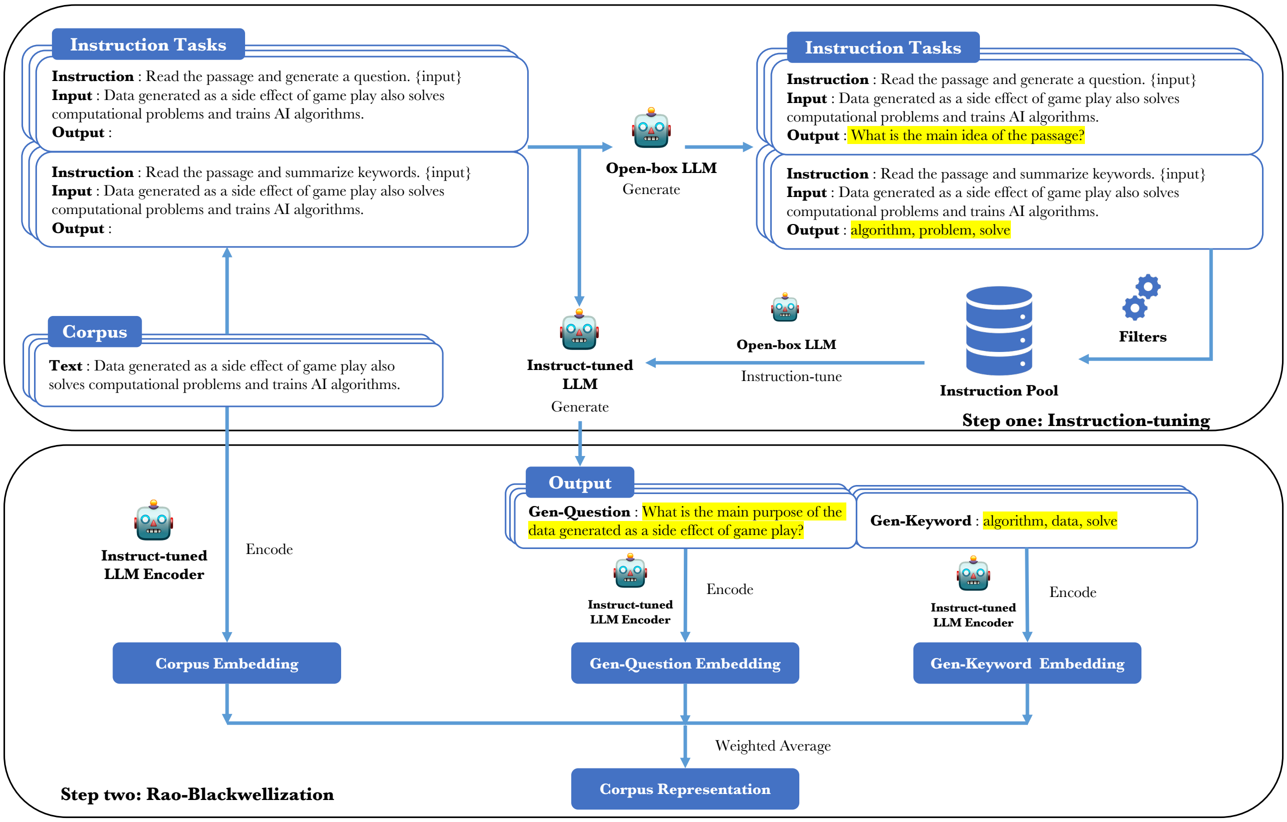
Instruction Tasks
Instruction : Read the passage and generate a question. {input}
Input : Data generated as a side effect of game play also solves computational problems and trains AI algorithms.
Output :
Instruction : Read the passage and summarize keywords. {input}
Input : Data generated as a side effect of game play also solves computational problems and trains AI algorithms.
Output :
Open-box LLM
Generate
Instruction Tasks
Instruction : Read the passage and generate a question. {input}
Input : Data generated as a side effect of game play also solves computational problems and trains AI algorithms.
Output : What is the main idea of the passage?
Instruction : Read the passage and summarize keywords. {input}
Input : Data generated as a side effect of game play also solves computational problems and trains AI algorithms.
Output : algorithm, problem, solve
Corpus
Text : Data generated as a side effect of game play also solves computational problems and trains AI algorithms.
Instruct-tuned LLM
Generate
Open-box LLM
Instruction-tune
Instruction Pool
Filters
Step one: Instruction-tuning
Output
Gen-Question : What is the main purpose of the data generated as a side effect of game play?
Gen-Keyword : algorithm, data, solve
Instruct-tuned LLM Encoder
Encode
Instruct-tuned LLM Encoder
Encode
Instruct-tuned LLM Encoder
Encode
Corpus Embedding
Gen-Question Embedding
Gen-Keyword Embedding
Weighted Average
Corpus Representation
Step two: Rao-Blackwellization

Figure 2: A high-level overview of Encoder-Decoder corpus representation. In the first instruction-tuning step, given a set of instruction tasks (in our case keyword summarization: "Read the passage and summarize keywords." and question generation: "Read the passage and generate a question."), the open-box LLM will generate instruction following examples which are passed through filters for quality control. The filtered examples form an instruction pool and are used to instruction-tune the open-box LLM. In the second Rao-Blackwellization step, by prompting the instruct-tuned LLM using the same instructions as in the first step, synthetic questions and keywords are generated for the corpus. Both the corpus and the generated sequences are encoded by the LLM encoder and the weighted average of their embedding is used as corpus representation.

can be `Harry Potter and the Philosopher's Stone`, whereas C_1 is `Harry Potter and the Sorcerer's Stone`. $Q_i = Q_{i1}, ..., Q_{im}$

Given a pre-trained encoder-decoder LLM, besides treating the encoder as a text representation model, we consider it as a random variable, where the sample space consists of the range of the possible embedding values, and the corresponding probability measure to each text portion.

$$\texttt{Encoder}(\cdot) : \textit{text} \mapsto \textit{embedding} \tag{1}$$

where the embedding refers to the sentence embedding of the text.

We assume that an effective encoder maps each group of queries Q_i near a group center in the high-dimensional embedding space and also maps the corresponding C_i to the surrounding area so that Q_i and C_i are well associated. For example, when we have $Q_{21} \in Q_2$ query, the retrieval system will retrieve the C_2 corpora which is the closest to the query (Figure 1).

Corpus Embedding as an Expectation Estimator The group center is a comprehensive depiction of the entire group and is indicative to distinguish from other groups. With the pre-trained $\texttt{Encoder}(\cdot)$, the group center is essentially the expected value of the embedding of each group's queries, denoted by $\mathbb{E}(\texttt{Encoder}(Q_i))$. When we use the embedding of the corpus, i.e. $\texttt{Encoder}(C_i)$, as its representation, we are using it to estimate the group center $\mathbb{E}(\texttt{Encoder}(Q_i))$. This is effective when we don't have any information from the query group.

Application of the Rao–Blackwell theorem Assume we have relevant queries Q_{i1},Q_{i2}, ..., Q_{im} for corpus C_i. Then $\frac{1}{m}\sum_{j=1}^{m}$ $\texttt{Encoder}(Q_{ij})$ is a sufficient statistics to estimate $\mathbb{E}(\texttt{Encoder}(Q_i))$.

According to Rao–Blackwell Theorem: If $g(\mathbf{X})$ is any kind of estimator of a parameter θ, then the conditional expectation of $g(\mathbf{X})$ given $T(\mathbf{X})$, namely $\mathbb{E}(g(x)|T(x))$, where T is a sufficient statistic, is typically a better estimator of θ, and is never worse. Plug in Equation (2), we get an improved estimator for $\mathbb{E}(\texttt{Encoder}(Q_i))$, which is $\mathbb{E}(\texttt{Encoder}(C_i)|\frac{1}{m}\sum_{j=1}^{m}\texttt{Encoder}(Q_{ij}))$.

$$\begin{aligned} g(x) &= \texttt{Encoder}(C_i) \\ T(x) &= \frac{1}{m}\sum_{j=1}^{m}\texttt{Encoder}(Q_{ij}) \\ \theta &= \mathbb{E}(\texttt{Encoder}(Q_i)) \end{aligned} \tag{2}$$

With some regularity assumptions, e.g., $C_i \in Q_i$ and $C_i = Q_{i1}$, the conditional expectation can be written as

$$\begin{aligned} &\mathbb{E}(\texttt{Encoder}(C_i)|\frac{1}{m}\sum_{j=1}^{m}\texttt{Encoder}(Q_{ij})) \\ &= \frac{1}{m}\sum_{j=1}^{m}\texttt{Encoder}(Q_{ij}) \\ &= \frac{1}{m}\texttt{Encoder}(C_i) + \frac{1}{m}\sum_{j=2}^{m}\texttt{Encoder}(Q_{ij}) \end{aligned} \tag{3}$$

We expect to achieve better performance with this formula for corpus representation. An intuitive understanding is that it gets closer to the relevant queries' embedding in the vector space (Figure 1).

3.2 Synthetic Query Generation

Obtaining a comprehensive set of labeled queries is labor-intensive and costly, especially in low resource setting. LLMs have built their reputation as generative AI models and are capable of following well designed instructions. Not only can the model generate text, but it also can output the generation probability of the text. We denote the generation model by $\texttt{LLM}(\cdot)$, then the generation can be written as,

$$\hat{Q}_{ij}, \hat{P}(\hat{Q}_{ij}) = \texttt{LLM}(\texttt{Instruction} + C_i) \tag{4}$$

where $\hat{Q}_{ij}$ is the generated query and $\hat{P}(\hat{Q}_{ij})$ is the generation probability. The instruction is a pre-defined generation task, for example "write a question for" or "what are the keywords of".

3.3 Corpus Representation

Plug in the generated synthetic queries, let $R(C_i)$ denote the final representation of corpora C_i, the Equation (3) becomes a weighted average of original corpora embedding and its synthetic query embedding,

$$\begin{aligned} R(C_i) \hat{=}\ & w_0\texttt{Encoder}(C_i) + \\ & (1 - w_0)\sum_j \hat{P}(\hat{Q}_{ij})\texttt{Encoder}(\hat{Q}_{ij}) \end{aligned} \tag{5}$$

where w_0 is a hyper-parameter that is tuned on a subset of test queries. Equation (5) is our proposed corpus representation for the dual-encoder retrieval system. Note that we can generate different types of synthetic queries in Equation (4) using various instructions, and we can generate multiple sequences for each instruction by adopting decoding strategies such as beam search. We can also improve the quality of the generated queries through instruction-tuning as follows.

3.4 Instruction-Tuning the LLM

While LLM demonstrates reasonable text generation capabilities, its ability to follow specific instructions can be honed. Instruction-tuning focuses on training a model to precisely follow the provided instructions.

As we don't have the query-corpora labeled data, we propose to self-instructed-tuning the LLM on its self-generated quality (i.e. gated) responses following given instructions to enhance synthetic queries generation relevance. This approach has demonstrated its effectiveness [31]. The instruct-tuned LLM is then used to prepare the synthetic queries for the corpus representation augmentation as in Equation (6).

$$\begin{aligned} &\hat{Q}_{ij}, \hat{P}(\hat{Q}_{ij}) = \\ &\texttt{InstructTunedLLM}(\texttt{Instruction} + C_i) \end{aligned} \tag{6}$$

We use the same instructions across the entire framework, including generation and training. Figure 1 shows a schematic diagram that although the generated queries from an open-box pre-trained LLM may not effectively enrich the corpora, after instruction-tuning, the generated synthetic queries become more relevant and the corpus representation can be improved consequently.

4 Experiments

4.1 Datasets

In this work, we tested four IR datasets where the summary of the database is shown in Table 1. ***English:*** (1) NFCorpus [1] has automatically extracted relevance judgments for medical documents. (2) SciFact [30] consists of expert-annotated scientific claims with abstracts and rationales. (3) SCIDOCS [3] has seven document-level tasks from citation prediction, document classification, and recommendation. ***German:*** (4) GermanQuAD [19] has the relevant information for high complex German QA with a large size of corpora. Due to computation resource limits, we downsample the corpus in SCIDOCS and GermanQuAD datasets, where we ensure the downsampled corpus include all relevant corpus for test queries. Note that such downsampling does not prevent us from fairly comparing the zero-shot retrieval efficacy of our approach with open-box LLMs because all experiments are performed under the same data setting. To help the encoder capture the fine-grained semantic interaction between queries and corpus, we divide each corpora into multiple sentences using the PunktSentenceTokenizer [1] from nltk package and use the sentence level multi representation for the corpora, meaning that if any of the sentence is retrieved, the passage is retrieved.

4.2 Baseline

We compare between the corpus-only representation and our proposed augmented corpus representation in zero-shot experiments under the dual-encoder framework. To obtain the representation of a sequence from the encoder, we perform mean aggregation over the last hidden state of each token following [21]. We measure the relevance between query and corpus using cosine similarity.

[1] https://www.nltk.org/api/nltk.tokenize.PunktSentenceTokenizer.html

Table 1: Details of datasets used. The size of corpus is downsampled to 15K in SCIDOCS and 10K in GermanQuAD. Filtered Queries: Generated synthetic queries from FLAN-T5-Large with filtering.

Dataset	Language	Test Queries	Corpus	Filtered Queries
NFCorpus	English	323	3.6K	5.9K
SciFact	English	300	5.1K	8.2K
SCIDOCS	English	1K	25.6K	29.4K
GermanQuAD	German	2K	2.8M	17.5K

Table 2: Open-box encoder-only average performance for passage level and sentence level indexing. Model is FLAN-T5. ♠: NDCG@10. ♣: MRR@100 ♡: Recall@100.

Index	Metric Model	♠	♣	♡
Passage	Base	5.79	8.02	22.75
	Large	8.78	10.63	32.43
Sentence	Base	22.02	25.96	43.54
	Large	**23.15**	**26.53**	**46.18**

To understand the superiority of our approach, we compare with three different SOTA models: (1) mDPR [34, 35] is a variation of DPR model [12] which replaces BERT to multilingual BERT [7] to support non-English languages for retrieval tasks. (2) T-Systems [28] is developed for computing sentenc embeddings for English and German texts. It uses a XLM-RoBERTa [4] and is fine-tuned with English-German datasets. (3) mBART-Large [29] is a multilingual Sequence-to-Sequence generation model. It supports 50 languages and we consider it for comparison in same model structure (i.e. encoder-decoder). Lastly, we compare with docTTTTTquery [23] to understand the effectiveness of our corpus representation augmentation.

4.3 Encoder-Decoder Models

T5 is an encoder-decoder model pre-trained on a combination of unsupervised and supervised tasks, where each task is transformed into a text-to-text format [26]. FLAN-T5 is an enhanced version of T5 fine-tuned on a mixture of tasks [32]. Considering that these types of models are open source, offer various sizes, support English and German, and have an encoder-decoder architecture, we leverage the FLAN-T5-Base and Large models in our experiments.

4.4 Instruction Query Generation

For instruction query generation and instruction-tuning, we consider two types of instructions (i.e. keyword summarization and question generation) as shown in Figure 2. We also develop a filter to improve the quality of generated instructions. If the task is keyword summarization, the number of keywords should be smaller than the half number of sentences in corpus. If it's question generation, the generated sequence should end with a question mark. The filter is simple, leaving room for further improvement. The numbers of the filtered synthetic queries are shown in Table 1.

4.5 Hyperparameter Setting

When performing instruction-tuning, we use the same hyperparameter setting for all the models. Specifically, we use the AdaFactor optimizer with learning rate 0.0001, batch size 16, and the number of epochs 30. Early stopping is performed when the validation loss shows no improvement for five consecutive epochs.

When generating queries using FLAN-T5 models, we only consider one returned sequence for each instruction and assume they are equally important. We denote the generated question and keywords as $\hat{question_i}$ and $\hat{keywords_i}$. We tested the multiple weighting methods for corpus representation where the best approach is giving the weight on the original corpus as $w_0 = 0.6$, so that each of $\hat{question_i}$ and $\hat{keywords_i}$ has the weight 0.2. Thus, the corpus representation is:

$$\begin{aligned}&\mathtt{R}(C_i) = 0.6 \times \mathtt{Encoder}(C_i) + 0.2 \times \\ &(\mathtt{Encoder}(\hat{question_i}) + \mathtt{Encoder}(\hat{keywords_i}))\end{aligned} \tag{7}$$

5 Results and Discussion

5.1 Corpora vs Sentence Indexing

We evaluate whether the sentence level multi-representation can capture the semantic interaction between the corpora and the query. Results for FLAN-T5 models using encoder-only representation are shown in Table 2. The sentence level multi-representation embedding technique outperforms the corpora level single representation by a large margin across all datasets. As the model size increases, the performance also gets better. Note that our approach uses no labeled data to achieve on par performance as SOTA models, and sentence level indexing is a way we do for chunking. According to the promising empirical results, we will apply the sentence level multi-representation technique to all the following experiments.

5.2 Overall Results

Table 3 describes the performance of FLAN-T5 models regarding instruction-tuning. Overall, we can mostly find the improvements of performances in all metrics after instruction-tuning, except for SCIDOCS. This is mainly because the quality of generated queries after instruction-tuning are proper and detailed (Table 6), and also each synthetic query is less overlapped which makes the corpora distinguishable. The influence of instruction-tuning is greater in larger model since it can have better generation capability and be more affected by fine-tuning with instructions.

Table 4 compares our approach and SOTA models in zero-shot scenarios. To clarify, FLAN-T5-Base has similar size as other SOTA models which can be considered as a fair comparison. First of all, instruct-tuned FLAN-T5-Base shows the boosted averaged results than other SOTA models which reveals the prowess of our approach. Considering a larger model (i.e. Tuned-FLAN-T5-Large) enhances the performance further. Thus, our suggested approach is consistently applicable in different size of models where the larger one promises the better performances.

5.3 Ablation Study

Optimal Corpus Representation From our findings, new corpus representation based on synthetic queries from instructions is useful to improve the retrieval performances. To define the optimal weights

Table 3: Comparison of performances according to instruction-tuning.

Instruction-tuning	Dataset	NFCorpus			SciFact			SCIDOCS			GermanQuAD			**Avg**		
	Model \ **Metric**	♠	♣	♡	♠	♣	♡	♠	♣	♡	♠	♣	♡	♠	♣	♡
No	Base	12.15	26.58	15.8	29.62	28.54	66.28	6.4	13.39	17.74	49.41	45.82	83.17	24.39	28.58	45.75
	Large	10.42	23.41	14.6	30.66	28.84	71.45	**7.22**	14.07	22.1	50.82	47.24	83.61	24.78	28.39	47.94
Yes	Base	**12.3**	27.02	**16.24**	30.72	29.59	65.06	6.04	12.7	16.5	52.39	48.54	84.44	25.36	29.46	45.56
	Large	11.91	**27.04**	15.85	**32.03**	**29.92**	**73.21**	7.16	**14.57**	**22.39**	**55.49**	**51.99**	**86.79**	**26.65**	**30.88**	**49.56**

Table 4: Comparison with SOTA. Tuned: Model with instruction-tuning.

Dataset		NFCorpus			SciFact			SCIDOCS			GermanQuAD			**Avg**		
Model	**Size** \ **Metric**	♠	♣	♡	♠	♣	♡	♠	♣	♡	♠	♣	♡	♠	♣	♡
mDPR	177M	8.30	19.19	11.57	23.46	21.87	58.94	4.75	10.26	15.98	**57.09**	**53.19**	**87.67**	23.40	26.13	43.54
T-Systems	278M	**15.32**	**29.14**	**17.05**	25.32	23.74	59.29	**8.38**	**17.64**	**23.82**	33.93	30.95	64.14	20.74	25.37	41.08
mBART-Large	331M	1.87	5.87	4.56	23.85	22.52	52.53	3.58	7.82	12.69	34.06	31.47	63.31	15.84	16.92	33.27
Tuned-FLAN-T5-Base	109M	12.30	27.02	16.24	**30.72**	**29.59**	**65.06**	6.04	12.70	16.50	52.39	48.54	84.44	**25.36**	**29.46**	**45.56**
Tuned-FLAN-T5-Large	341M	11.91	27.04	15.85	***32.03***	***29.92***	***73.21***	7.16	14.57	22.39	55.49	51.99	86.79	***26.65***	***30.88***	***49.56***

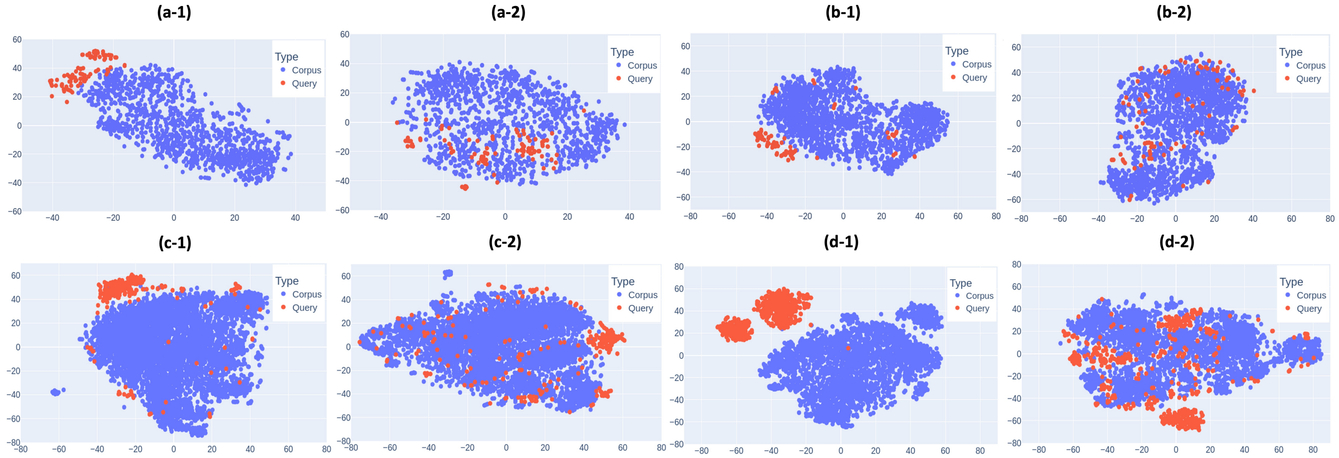

Figure 3: t-SNE distributions for corpus representation generated from FLAN-T5-Large. (a-d) NFCorpus, SciFact, SCIDOCS, GermanQuAD. (1-2) Original corpus, Weighted corpus with synthetic queries after instruction-tuning.

in corpus representation, we investigate the four different weight methods: (1) Equal: Giving the equal weights for corpus and generated synthetic queries (i.e. keyword, question). (2) Manual: It is same as Equation (7). (3) BERTScore: Giving the weights based on BERTScore (F1) with BERT-Base-Multilingual-Cased model [6]. Equation (8) shows the details of it. (4) BERTScore$_{Softmax}$: Similar as BERTScore but including the Softmax.

$$
\begin{aligned}
&X : \widehat{keywords}_i, \widehat{question}_i \\
&\mathtt{BERT}_X : \mathtt{BERTScore} \text{ between } X \text{ and } C_i \\
&\mathtt{Denominator} = 1 + \mathtt{Sum}(\mathtt{BERT}_X) \\
&\mathtt{Weight}_X = \frac{\mathtt{BERT}_X}{\mathtt{Denominator}}, \\
&\mathtt{Weight}_{C_i} = \frac{1}{\mathtt{Denominator}}
\end{aligned}
\tag{8}
$$

Table 5 shows the overall performances according to the different weight approaches in corpus representation. First of all, the equal weight approach shows the worst performance which confirms that the corpus basically contains the most relevant information for queries which should be weighted more. Also, extracted keywords and questions mostly have the essential contexts but partial information of corpus which is not enough to include the semantic meaning of corpus. Thus, manual weighting with emphasis on corpus promises the better result than BERTScore approaches.

Effectiveness of Instruction-tuning Table 6 gives the examples of generated synthetic queries. In keyword summarization, open-box extracts the ambiguous and meaningless words (first example) or a simple copy of sentence (second example) as keywords while instruction-tuning helps to observe the whole corpus to extract the core keywords. For question generation, open-box generates the general (third example) or unanswerable questions (fourth example)

Table 5: Example of synthetic queries according to the instruction-tuning. FLAN-T5-Large is used for generating the examples.

Corpus	Instruction Type	Open-box	Instruct-tuned
Semantic Space is a pervasive computing infrastructure that exploits semantic Web technologies to support explicit representation, expressive querying, and flexible reasoning of contexts in smart spaces.	Keyword	context, support, query, semantic, space	semantic space
Fluorometric titration of E. coli single-stranded DNA binding protein with various RNAs showed that the protein specifically and cooperatively binds to its own mRNA. The binding inhibited in vitro expression of ssb and bla but not nusA. This inhibition takes place at a physiological concentration of SSB. The function of the protein in gene regulation is discussed.	Keyword	The single-stranded DNA binding protein(SSB) specifically and cooperatively binds to its own mRNA.	mRNA, protein, titration
This paper describes an aggregation and correlation algorithm used in the design and implementation of an intrusion-detection console built on top of the Tivoli Enterprise Console (TEC). The aggregation and correlation algorithm aims at acquiring intrusion-detection alerts and relating them together to expose a more condensed view of the security issues raised by intrusion-detection systems.	Question	What is the purpose of the paper?	What is the purpose of the aggregation and correlation algorithm?
ESC is to create an inventory of cardiovascular disease registries and a task force on data standardization	Question	What is the purpose of the task force?	What is the purpose of the ESC?

Table 6: Different weight methods for corpus representation. Model is based on FLAN-T5.

Corpus Weights	Metric Model	♠	♣	♡
N/A	Base	22.02	25.96	43.54
	Large	23.15	26.53	46.18
Equal	Base	18.25	21.98	38.76
	Large	17.92	21.60	39.93
Manual	Base	24.39	**28.58**	45.75
	Large	**24.78**	28.39	**47.94**
BERTScore	Base	22.35	26.12	43.58
	Large	21.97	25.53	45.2
BERTScore$_{Softmax}$	Base	20.13	23.64	40.68
	Large	19.52	23.12	42.72

Table 7: Different corpus representation augmentation. Model is based on FLAN-T5. Note that we evaluated on English datasets.

Model	Metric Method	♠	♣	♡
Base	docTTTTTquery	4.94	10.12	14.33
	Our approach	**16.35**	**23.1**	**32.6**
Large	docTTTTTquery	6.87	11.81	19.93
	Our approach	**17.03**	**23.84**	**37.15**

while instruction-tuning gives the detailed and suitable questions which can be accountable by the specific corpus.

Figure 3 shows the distributions of embeddings of corpora and test queries based on FLAN-T5-Large. Overall, the weighted corpus representation and instruction-tuning spread out the corpora embeddings to make them distinguishable. Also, it helps to locate the test queries closer to the corpora. Thus, our approach helps to integrate the crucial and detailed synthetic queries for corpus representation which helps to generate the unique corpora to achieve the enhanced retrieval performances.

Effectiveness of Corpus Representation Augmentation We compare with other corpus representation augmentation, docTTTTTquery [23], to validate our corpus augmentation. Here, we follow the default strategy of docTTTTTquery: top-10 with 40 predictions appending on corpus. According to Table 7, we demonstrate significant improvement via our approach - embedding level augmentation with representations from self-instructed-tuned model. Based on this finding, we can confirm that augmenting representation on embedding level is more effective than on input text level with concatenation as docTTTTTquery, and our self-instructed-tuned model performs better than their supervised representation generation model.

6 Conclusion

In our research, we propose the unsupervised text representation learning technique through self-instructed-tuning encoder-decoder LLMs. Based on the Rao-Blackwell theorem, we leverage the embeddings of synthetically generated queries (i.e. questions and keywords) to augment the corpus representation for the dual-encoder retrieval framework. In zero-shot experiments, our proposed corpus representation consistently improves the performance over encoder-only corpus representation. Even if the open-box LLM was not pre-trained on retrieval task and there is no labeled modeling data, after fine-tuning with our approach it exceeds the SOTA models across different datasets, presenting the high effectiveness and data efficiency of our method in retrieval tasks.

In future work, we plan to explore our proposed method on separate encoder and decoder models.

References

[1] Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A Full-Text Learning to Rank Dataset for Medical Information Retrieval. *Proceedings of the 38th European Conference on Information Retrieval*. http://www.cl.uni-heidelberg.de/~riezler/publications/papers/ECIR2016.pdf

[2] Sukmin Cho, Soyeong Jeong, Wonsuk Yang, and Jong C. Park. 2022. Query Generation with External Knowledge for Dense Retrieval. 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*, Eneko Agirre, Marianna Apidianaki, and Ivan Vulic (Eds.). Association for Computational Linguistics, 22–32. https://doi.org/10.18653/v1/2022.deelio-1.3

[3] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2270–2282. https://doi.org/10.18653/v1/2020.acl-main.207

[4] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale. *CoRR* abs/1911.02116 (2019). arXiv:1911.02116 http://arxiv.org/abs/1911.02116

[5] Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. 2023. Promptagator: Few-shot Dense Retrieval From 8 Examples. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/pdf?id=gmL46YMpu2J

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv* abs/1810.04805 (2019). https://api.semanticscholar.org/CorpusID:52967399

[8] Zhe Dong, Jianmo Ni, Dan Bikel, Enrique Alfonseca, Yuan Wang, Chen Qu, and Imed Zitouni. 2022. Exploring Dual Encoder Architectures for Question Answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 9414–9419. https://aclanthology.org/2022.emnlp-main.640

[9] Daniel Gillick, Alessandro Presta, and Gaurav Singh Tomar. 2018. End-to-End Retrieval in Continuous Space. *ArXiv* abs/1811.08008 (2018). arXiv:1811.08008 http://arxiv.org/abs/1811.08008

[10] Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskénazi, and Jeffrey P. Bigham. 2022. InstructDial: Improving Zero and Few-shot Generalization in Dialogue through Instruction Tuning. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:249062857

[11] Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring. In *International Conference on Learning Representations*. https://api.semanticscholar.org/CorpusID:210063976

[12] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

[13] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

[14] O. Khattab and Matei A. Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020). https://api.semanticscholar.org/CorpusID:216553223

[15] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: a Benchmark for Question Answering Research. *Trans. Assoc. Comput. Linguistics* 7 (2019), 452–466. https://doi.org/10.1162/tacl_a_00276

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv* abs/1907.11692 (2019). https://api.semanticscholar.org/CorpusID:198953378

[17] Y Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2020. Sparse, Dense, and Attentional Representations for Text Retrieval. *Transactions of the Association for Computational Linguistics* 9 (2020), 329–345. https://api.semanticscholar.org/CorpusID:218470027

[18] Ji Ma, Ivan Korotkov, Yinfei Yang, Keith B. Hall, and Ryan T. McDonald. 2021. Zero-shot Neural Passage Retrieval via Domain-targeted Synthetic Question Generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, 1075–1088. https://doi.org/10.18653/v1/2021.eacl-main.92

[19] Timo Möller, Julian Risch, and Malte Pietsch. 2021. GermanQuAD and GermanDPR: Improving Non-English Question Answering and Passage Retrieval. *ArXiv* abs/2104.12741 (2021). arXiv:2104.12741 https://arxiv.org/abs/2104.12741

[20] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016 (CEUR Workshop Proceedings, Vol. 1773)*, Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[21] Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 1864–1874. https://doi.org/10.18653/v1/2022.findings-acl.146

[22] Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large Dual Encoders Are Generalizable Retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 9844–9855. https://aclanthology.org/2022.emnlp-main.669

[23] Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019. From doc2query to docTTTTTquery. *Online preprint* 6 (2019), 2.

[24] Rodrigo Frassetto Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document Expansion by Query Prediction. *ArXiv* abs/1904.08375 (2019). arXiv:1904.08375 http://arxiv.org/abs/1904.08375

[25] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv* abs/2203.02155 (2022). https://api.semanticscholar.org/CorpusID:246426909

[26] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[27] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:201646309

[28] T-Systems. 2020. T-System Model. https://huggingface.co/T-Systems-onsite/cross-en-de-roberta-sentence-transformer

[29] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning. (2020). arXiv:2008.00401 [cs.CL]

[30] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 7534–7550. https://doi.org/10.18653/v1/2020.emnlp-main.609

[31] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. 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*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 13484–13508. https://aclanthology.org/2023.acl-long.754

[32] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned Language Models are Zero-Shot Learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. https://openreview.net/forum?id=gEZrGCozdqR

[33] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernández Ábrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual Universal Sentence Encoder for Semantic Retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, Asli Celikyilmaz and Tsung-Hsien Wen (Eds.). Association for Computational Linguistics, 87–94. https://doi.org/10.18653/v1/2020.acl-demos.12

[34] Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A Multi-lingual Benchmark for Dense Retrieval. *arXiv:2108.08787* (2021).

[35] Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. 2022. Towards Best Practices for Training Multilingual Dense Retrieval Models. arXiv:2204.02363 [cs.IR]

[36] Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2022. Dense Text Retrieval based on Pretrained Language Models: A Survey. *ArXiv* abs/2211.14876 (2022). https://doi.org/10.48550/arXiv.2211.14876 arXiv:2211.14876