# Stop Generating Simple Question as a Query!

**Anonymous ACL submission**

## Abstract

A prevailing strategy for zero-shot retrieval entails the construction of synthetic queries from documents. However, these generated queries tend to be simple and concise, hence falling short in adequately representing diverse retrieval tasks. An alternative approach harnesses the capability of large language models (LLMs) for in-context learning, enabling the retriever to adapt effectively to the target domain. Nonetheless, such endeavours to discover the unspecified intents demand massive computational resources. In this paper, we challenge the conventional approach of creating simple questions as queries. We propose **TOPiC**, which directly generates `task-oriented` queries. TOPiC achieves the highest performance on 6 non-QA datasets, as well as second on entire BeIR benchmark. Our study underscores the potential benefits of incorporating stylistic elements into the query generation procedure.

## 1 Introduction

Information retrieval has significantly facilitated the process of locating relevant documents in response to user requests. With the advent of dense retrieval (Karpukhin et al., 2020), a substantial body of research has concentrated on the supervised alignment of the latent spaces within query and passage encoders (Gao and Callan, 2021; Ni et al., 2021). However, this might be impractical in real-world scenarios due to the high costs associated with collecting labeled data across numerous domains. Accordingly, a training framework which doesn't rely on the availability of relevance labels, referred to as *zero-shot*, has been widely explored (Izacard et al., 2021; Santhanam et al., 2021).

A common zero-shot approach involves the creation of pseudo-queries from a specific target corpus. MSMARCO (Campos et al., 2016), an extensive question answering (QA) dataset, is a frequent choice for training the query generator (Cheriton, 2019; Ma et al., 2021). This process relies on the belief that the intrinsic relationship within MSMARCO can be effectively transferred to downstream tasks (Dai et al., 2022). However, the distinct characteristics of tasks, including the domain, intent, and query style, present variations across different datasets. For instance, MSMARCO queries typically assume the form of brief questions seeking for an entity, whereas Scifact (Wadden et al., 2020) focuses on retrieving factual claim within scientific domain, which we define as a *non-QA* task. Meanwhile, prompting LLM to generate a '`query`' still faces limitations attributed to the general nature of questions. This discrepancy in task styles accentuates the necessity for query generation approaches, tailored to target domains.

Recent efforts have focused on the integration of task-specific intentions into various retrieval tasks (Asai et al., 2022; Hashemi et al., 2023). Promptagator (Dai et al., 2022), in particular, has gained attention for its remarkable performance. Few-shot examples are extracted to perform in-context learning (Brown et al., 2020), utilizing LLM with a size of 137B, to comprehend and better align the stylistic attributes of the target domain. Nevertheless, this employment comes at the cost of intensive computational requirements (Brown et al., 2020) to capture the latent intents.

In this study, we introduce ***T**ask-**O**riented **P**seudo-Que**r**i**e**s **C**onstruction* (TOPiC), which is designed to enhance the query construction process by directly incorporating concise task-specific descriptions (Figure 1). This novel approach facilitates the queries to be finely attuned to the unique characteristics of each task. TOPiC improves over 1.3 nDCG on *non-QA* tasks, highlighting its efficacy in domain adaptation. Moreover, our method achieves second rank overall, while upholding efficiency through the exclusion of few-shot examples, enabling the utilization of 45x smaller query generator. Our empirical findings offer valuable insights

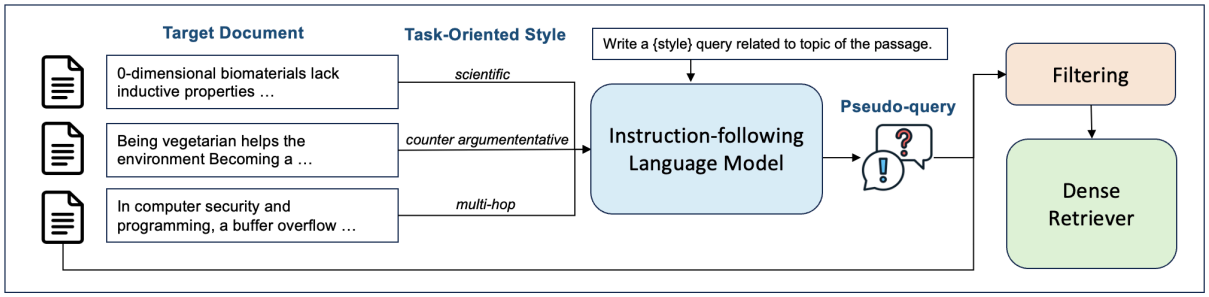


Figure 1: Overview of our training pipeline. First, documents are fed into instruction-following language model with their task-oriented style descriptions in the prompt. After queries are generated, together with the documents they form a synthetic dataset. After applying filtering mechanism based on cosine similarity for removing low-quality pairs, the dataset can be applied to any existing retrieval methods for training.

into the development of synthetic datasets, tailored to the unique demands of specific retrieval tasks.

## 2 Methodology

### 2.1 Task-oriented Query Generation

In the process of synthesizing task-oriented queries, we employ an instruction-following language model. This model receives a concatenated passage and a predefined prompt as an input. We universally employ the following prompt: '`Write a {style} query related to topic of the passage. Do not directly use wordings from the passage. passage`'. This serves to channel the focus of the generator towards task-specific topic. The stylistic attributes, illustrated in Figure 2, are mainly sourced from BeIR paper (Thakur et al., 2021a). However, for 3 datasets[1], where the provided descriptions were inappropriate to coordinate the term '`query`', we refer to the abstract of each paper. It is noteworthy that we have encouraged the avoidance of directly referencing the passage, as doing so might act as an easy shortcut during generation.

### 2.2 Cosine Filtering

Generating sentences from lengthy passages can yield queries of poor quality, which have the potential to disrupt the training process. While Dai et al. (2022) has demonstrated the efficacy of consistency filtering by enforcing a round-trip consistency (Alberti et al., 2019), it requires the retrieval of entire queries, which has the complexity of $O(QD)$. Instead, initial experiments on MSMARCO show that irrelevant queries tend to exhibit low cosine similarity scores to their passages (Table 3). An initial retriever is trained for a single epoch on the synthetic dataset, followed by a filtration mechanism designed to exclude queries that fall below a certain threshold, with a complexity of $O(Q)$.

[1]FEVER (Thorne et al., 2018), Scidocs (Cohan et al., 2020), NFCorpus(Boteva et al., 2016)

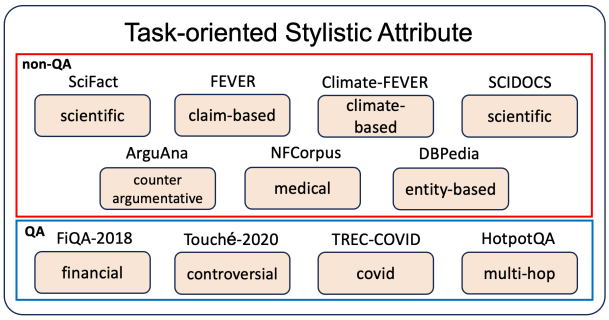


Figure 2: Task-oriented stylistic attributes used in instruction for query generation. Representative attributes are derived from BeIR and original dataset papers.

### 2.3 Dense Retrieval

Our methodology demonstrates compatibility with previously established retrievers, including DPR (Karpukhin et al., 2020), GTR (Ni et al., 2021), and others. In particular, our methodology adopts the training framework of DPR as baseline, which finetunes BERT (Devlin et al., 2019) encoder via in-batch negative samples of synthetic data.

### 2.4 Dataset

BeIR is a comprehensive benchmark for zero-shot retrieval, encompassing 18 tasks. From 14 publicly available datasets, we focus on 11 of them, following Dai et al. (2022) for a direct comparison. Since task-oriented query generation may yield more advantages if the queries are not in question format, we categorize the datasets into two groups. The *non-QA* group, consisting of 7 datasets, involves queries that seek to provide information or facts rather than specific answers. The *QA* group consists of 4 datasets, which are mainly formed of questions. This division facilitates a meticulous examination across different types of tasks.

| | *Unsupervised* | | *Supervised* | | *Supervised+Query Generation* | | | | *TOPiC* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Promptagator | | | | |
| | BM25 | Contriever | TAS-B | GTR-XXL | GenQ | GPL | +Zero | +Few | Base | +DPR | +GPL |
| *non-QA datasets* | | | | | | | | | | | |
| scifact | 66.5 | 64.9 | 64.3 | 66.2 | 64.4 | 67.4 | 62.3 | 65.0 | 65.0 | 67.7 | **68.4** |
| fever | 75.3 | 68.2 | 70.0 | 74.0 | 66.9 | 75.9 | 76.2 | 77.0 | 61.2 | 67.9 | **78.8** |
| climate | 21.3 | 15.5 | 22.8 | **26.7** | 17.5 | 23.5 | 21.4 | 16.8 | 17.3 | 16.9 | 23.0 |
| scidocs | 15.8 | 14.9 | 14.9 | 16.1 | 14.3 | 16.9 | 16.3 | **18.5** | 16.7 | 17.9 | 16.9 |
| arguana | 31.5 | 37.9 | 42.9 | 54.0 | 49.3 | 55.7 | 53.8 | 59.4 | 55.9 | **61.2** | 59.3 |
| nfcorpus | 32.5 | 31.7 | 31.9 | 34.2 | 31.9 | **34.5** | 33.4 | 33.4 | 33.2 | 33.4 | 34.3 |
| dbpedia | 31.3 | 29.2 | 38.4 | **40.8** | 32.8 | 38.4 | 36.4 | 38.0 | 32.3 | 35.0 | 40.6 |
| **Avg.** | 39.2 | 37.5 | 40.7 | 44.6 | 39.6 | 44.6 | 42.8 | 44.0 | 40.2 | 42.9 | **45.9** |
| *QA datasets* | | | | | | | | | | | |
| fiqa | 23.6 | 24.5 | 30.0 | **46.7** | 30.8 | 34.4 | 40.4 | 46.2 | 31.2 | 32.7 | 33.9 |
| touché | **36.7** | 19.3 | 16.2 | 25.6 | 18.2 | 25.5 | 26.6 | 34.5 | 17.4 | 14.7 | 24.1 |
| trec-covid | 65.6 | 27.4 | 48.1 | 50.1 | 61.9 | 70.0 | 72.7 | **75.6** | 62.2 | 71.5 | 63.1 |
| hotpotqa | 60.3 | 48.1 | 58.4 | 59.9 | 53.4 | 58.2 | 60.4 | 61.4 | 53.9 | 55.5 | **61.6** |
| **Avg.** | 46.6 | 29.8 | 38.2 | 45.6 | 41.1 | 47.0 | 50.0 | **54.4** | 41.2 | 43.6 | 45.7 |
| **Total Avg.** | 41.9 | 34.7 | 39.8 | 44.9 | 40.1 | 45.5 | 45.4 | **47.8** | 40.6 | 43.1 | 45.8 |

Table 1: Model performances across BeIR datasets in nDCG@10. Our model with DPR outperforms GenQ and our baseline, revealing the limited efficacy of *question-style* queries in retrieval. When employed in conjunction with GPL, TOPiC achieves the highest average score among the *non-QA* datasets and ranks second across all tasks. Bold and underline indicates the best and second best score within each dataset, respectively.

# 3 Experiments

## 3.1 Implementation

We leverage FLAN-T5-XL (Chung et al., 2022), a language model capable of accommodating instructions. We mainly follow the hyperparameters from Dai et al. (2022) for query generation. While Promptagator-Few generates up to 8 queries from a maximum of 1M documents, we limit to 83K (Thakur et al., 2021b) to reduce training costs. Cosine filtering threshold at 0.25 was selected by experimenting on MSMARCO (Figure 4). Throughout the training phase, we employ DPR and GPL with DistilBERT-TAS-B (Hofstätter et al., 2021).

Among unsupervised models, we benchmark against BM25 and Contriever (Izacard et al., 2021). GenQ and GPL utilize MSMARCO-trained T5 (Raffel et al., 2019) generator and finetune TAS-B retriever. While GenQ adopts DPR, GPL enhances GenQ through negative mining and score distillation from a cross-encoder. GTR trains T5 encoder on large-scale QA datasets, despite not finetuning on the target tasks. Promptagator leverages 137B FLAN (Wei et al., 2021) as generator and GTR-base as retriever. Unlike Promptagator-Few, its zero-shot model only asks to generate '`query`'. Similarly, our baseline generates '`query`' without any stylistic features. We evaluate with nDCG@10 metric, a standard measure for BeIR. For further elaboration, please refer to Appendix A.

## 3.2 Main Results

For *non-QA* datasets, TOPiC+DPR slightly outperforms Promptagator-Zero, albeit in a more efficient manner. TOPiC+GPL exceeds other methods by 1.3 points, providing empirical support that using a simple question as a query may yield suboptimal results. Specifically, our method secures the highest scores in three datasets and performs comparably on others. GTR-XXL attains the top position on two datasets, despite using an large retriever, up to 44x the size of TAS-B. Interestingly, Promptagator fails to surpass others, indicating that its gains are primarily confined to *QA* datasets.

Regarding *QA* datasets, our method achieves the highest score on HotpotQA (Yang et al., 2018), while GPL and Promptagator exceed ours in the remaining domains. These outcomes suggest that factors beyond task-specific elements, such as transferring from large QA dataset or the use of LLM, play a pivotal role in *QA* tasks. Nonetheless, existing methods may not be familiar with multi-hop combinational queries in HotpotQA. As TOPiC considers such style while constructing queries, performance is further enhanced with GPL through mining and alignment of relevant documents.

# 4 Rethinking the Semantics of the '`Query`'

Historically, user-generated input in information retrieval systems has been designated as a '`query`', a terminology that dates back to foundational works

such as Luhn (1958). The etymological roots of 'query' suggest an interrogative nature, which could unintentionally introduce biases during query formulation process. This becomes especially pertinent when considering retrieval tasks where the desired input format does not strictly conform to the structure of a question. This paper aims to delve into the implications and potential reconfigurations when system-generated queries deviate from the conventional question-based format.

**`format-oriented query`** We replace '`{style} query`' in our prompt with '`{desired format}`'. The experimental results indicate varying effects over different tasks (Figure 3). FEVER and Scidocs exhibit improvements, while other two demonstrate a decline in scores. The nature of the tasks appears to play a significant role in these outcomes. It is challenging to define a task-specific style for FEVER, as it involves a `natural claim` generation from Wikipedia passages, which can vary widely in format. Scidocs also asks for an `article title` cited by a scientific article, where adherence to the desired title format may be beneficial.

As illustrated in Table 2, the enforcement of query generation in the desired format is feasible; however, it results a high degree of duplication. Conversely, the role of the term '`query`' within the prompt is to encourage diversity, as it focuses on various elements within the passage. The strong performance of TOPiC can be attributed to its harmonious approach, which inspires diversity and, concurrently, constrains the queries to be contextually relevant to the given task.

**Method Speciality** We generalize our findings to reveal the unique specialities of different approaches. While Promptagator performs better on most of *QA* datasets, we recommend our method on cases that demonstrates uniqueness, like HotpotQA. For *non-QA* datasets, TOPiC shows strong performance when task-specific style can be well defined. Although *format-oriented* datasets, such as FEVER, Scidocs, and NFCorpus don't have a single dominant method, we have showcased the potential of *format-oriented* queries, which may exhibit strong performance in such scenarios.

## 5 Related Work

Zero-shot retrieval has been mainly studied by Thakur et al. (2021a) along with their benchmark, BeIR, which comprises 18 multi-domain datasets

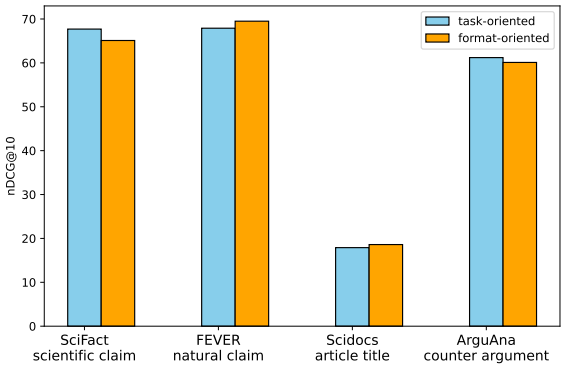


Figure 3: Comparison between utilization of task-oriented style and format-oriented style. Desired formats, described below the name of each dataset, are adopted from BeIR paper.

among 9 tasks. Our focus centers on the scenario where only the target corpus is available (Izacard et al., 2021; Ni et al., 2021; Gao et al., 2022). In response to this challenge, various methods have been explored to create synthetic labels by generating pseudo queries from the documents and jointly trained in a retriever. GenQ (Ma et al., 2021) utilizes MSMARCO to train T5 (Raffel et al., 2019) model in creating questions. Furthermore, GPL (Thakur et al., 2021b) performs negative mining of similar documents and distills cross-encoder score into a MarginMSE (Hofstätter et al., 2020) loss.

Another line of works made efforts to incorporate the underlying search intents to retrieval, such as prepending task-specific instruction to each query (Asai et al., 2022) or considering scenario where only human annotated target descriptions are available (Hashemi et al., 2023). In-context learning via demonstrating up to 8 few-shot examples can generate task-specific queries (Dai et al., 2022). However, it requires massive model size for significant improvement (Brown et al., 2020). Moreover, longer prompt accompanies higher computational cost since attention is quadratic.

## 6 Conclusion

In this work, we present TOPiC, a novel approach that overcomes the shortcomings of previous zero-shot retrieval methodologies. TOPiC mitigates issues associated with existing query generation methodologies, including reliance on simplistic questions and computational overheads of in-context learning with large language models. Our approach demonstrates superior performance and efficiency on BeIR datasets, indicating a promising direction for task-oriented query generation.

## Limitations

A particular task can exhibit diverse stylistic characteristics, while we adopted the most representative descriptions provided by the authors of each dataset. Careful selection of these variations through prompt engineering has the potential to augment performance, as exemplified by our ablation study focusing on *format-oriented* queries. Furthermore, our intentional utilization of FLAN-T5 is motivated by the usage of T5 in MSMARCO-based methods and the adoption of FLAN by Promptagator. We reserve the exploration of alternative language models, such as LLaMA (Touvron et al., 2023) and OPT (Zhang et al., 2022), to future investigations, seeking to discern their capacity to specialize in certain tasks.

## References


Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic qa corpora generation with roundtrip consistency. In *Annual Meeting of the Association for Computational Linguistics*.

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen tau Yih. 2022. Task-aware retrieval with instructions. In *Annual Meeting of the Association for Computational Linguistics*.

Vera Boteva, Demian Gholipour Ghalandari, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *European Conference on Information Retrieval*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Daniel Fernando Campos, Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, Li Deng, and Bhaskar Mitra. 2016. Ms marco: A human generated machine reading comprehension dataset. *ArXiv*, abs/1611.09268.

David R. Cheriton. 2019. From doc2query to doctttttquery.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. *ArXiv*, abs/2004.07180.

Zhuyun Dai, Vincent Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. *ArXiv*, abs/2209.11755.

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.

Thomas Diggelmann, Jordan L. Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *ArXiv*, abs/2012.00614.

Luyu Gao and Jamie Callan. 2021. Condenser: a pre-training architecture for dense retrieval. In *Conference on Empirical Methods in Natural Language Processing*.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise zero-shot dense retrieval without relevance labels. *ArXiv*, abs/2212.10496.

Helia Hashemi, Yong Zhuang, Sachith Sri Ram Kothur, Srivas Prasad, Edgar Meij, and W. Bruce Croft. 2023. Dense retrieval adaptation using target domain description. *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020. Improving efficient neural ranking models with cross-architecture knowledge distillation. *ArXiv*, abs/2010.02666.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy J. Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Hans Peter Luhn. 1958. A business intelligence system. *IBM J. Res. Dev.*, 2:314–319.

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 *Conference of the European Chapter of the Association for Computational Linguistics*.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2021. Large dual encoders are generalizable retrievers. *ArXiv*, abs/2112.07899.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Keshav Santhanam, O. Khattab, Jon Saad-Falcon, Christopher Potts, and Matei A. Zaharia. 2021. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *North American Chapter of the Association for Computational Linguistics*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021a. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021b. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *ArXiv*, abs/1803.05355.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

David Wadden, Kyle Lo, Lucy Lu Wang, Shanchuan Lin, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *ArXiv*, abs/2004.14974.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *ArXiv*, abs/2109.01652.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *ArXiv*, abs/2205.01068.

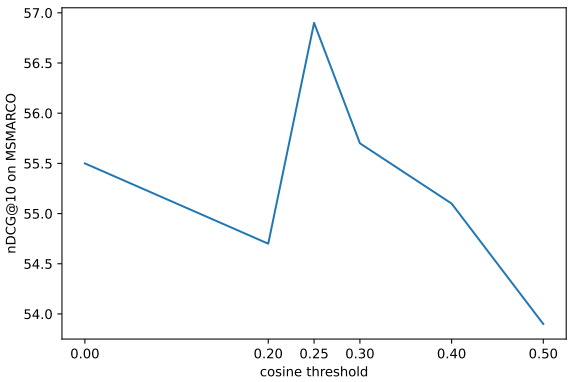


Figure 4: Performance on MSMARCO by differing cosine thresholds. Generated queries of cosine score lower than each value of x-axis is removed.

## A Implementation Details

We utilize the publicly available FLAN-T5-XL checkpoint[2], to generate 8 queries per passage. The hyperparameters for query generation are mainly derived from Dai et al. (2022). In instances where passages exceed 350 tokens, they are truncated, and query sampling is executed with a temperature of 1.0, employing parameters $k = 25$ and $p = 0.95$. We randomly sample 83K documents if the corpus size exceeds. For training the DistilBERT-TASB retriever, a batch size of 75 is adopted. If the corpus size is larger than 60K, a single epoch is conducted; otherwise, 3 epochs are performed. The training process incorporates a learning rate of 2e-5 and a warming step of 1000. In alignment with GPL's recommended configuration, we use 250K/|C| queries per passage when applying

[2]https://huggingface.co/google/flan-t5-xl

GPL, where |C| refers to the corpus size. For the prompt of our baseline, 'Read the passage and generate a query. passage' adopted from Promptagator-Zero, is used. All experiments are conducted on a single RTX 3090 GPU.

## B Cosine Filtering

Threshold value for cosine filtering is selected through experiments on MSMARCO, as shown in Figure 4. We further present examples and the corresponding scores for MSMARCO in Table 3.

## C Qualitative Analysis

In this section, we provide examples of 3 generated queries for each dataset. Notable distinction observed in *non-QA* datasets is that the model occasionally produces not only questions but also stylistic claims, effectively incorporating task-specific styles. For example, in ArguAna dataset, we observe that the generated queries adopt an opposing standpoint to its passage. In the case of *QA* datasets, the generated queries are well aligned with the topic of respective tasks. Moreover, we observe that the pseudo-queries of Touché2020 dataset vary in their structures. However, there are only 49 queries in the test split of Touché2020, with a predominant starting phrase of 'Should'. The mismatch in query style and small test split may have contributed to its relatively lower performance of TOPiC+DPR on this specific dataset.

In addition, the effectiveness of TOPiC may encounter challenges when the entire corpus does not uniformly encompass the task-oriented attributes. Climate-FEVER (Diggelmann et al., 2020), which comprised 7675 annotated claim-evidence pairs, targets climate-change retrieval under 5.42M Wikipedia corpus. Due to the limited presence of documents directly related to climate change and the fact that we only sample 1.5% of the entire corpus, the generated queries may exhibit anomalies. For example, a query like 'What is the mean sea level of Roy Taylor?' appears in Table 4, where Roy Taylor is a person. A possible mitigation strategy could involve a careful selection of passages that are related to climate, and subsequently utilizing this refined set to construct a synthetic dataset.

## D Format-oriented Query Examples

We present the examples of format-oriented queries from Section 4 in Table 2.

| Dataset | Generated query |
|---|---|
| Scifact | Cyclin A1 promotes meiosis in the mouse male testis and its absence causes sterility. |
| FEVER | Der Klassiker: Bayern v Borussia Dortmund football. |
| Scidocs | Executive attention and working memory: a review |
| ArguAna | Banning 3D printers could prevent them from forming an economically viable market |

Table 2: Examples of `format-oriented` queries. Queries are more alike to desired query styles.

| **passage** | |
|---|---|
| The two cells that come from the division of one cell are called daughter cells. odd, but the cells produced by cell division are called daughter cells, even in boys and men.) Each of the daughter cells needs to have a complete set of chromosomes. | |
| **generated query** | **cosine score** |
| How do you cite a website? | -0.14558335 |
| **passage** | |
| (As a result, media companies can post four to 10 times more often than brands and still see engagement.) An Edgerank Checker study posted on the Moz blog determined that one way to counteract the algorithm change might be to publish more frequently â as often as you have fresh, compelling content to share. | |
| **generated query** | **cosine score** |
| What will the internet do to me? | 0.25003907 |
| **passage** | |
| Gyarados is a massive threat whom one should be prepared for. A high base 125 Attack, an excellent offensive typing in Water, and a diverse movepool all make Gyarados a top-tier sweeper. | |
| **generated query** | **cosine score** |
| What is the best Pokemon to fight in Pokemon? | 0.50083715 |
| **passage** | |
| These nodules usually occur over exposed joints that are subject to trauma, such as the fingers joints and elbows, though occasionally they can occur elsewhere such as the back of the heel. They are usually non-tender and only occasionally painful and very rarely the overlying skin can become infected or even ulcerate. | |
| **generated query** | **cosine score** |
| What causes nodules in the hands? | 0.66780293 |
| **passage** | |
| If one gets Botox into the chin area to smooth out a pebbly chin, does the mentalis muscle eventually atrophy ?. Over what period of time would this... read more | |
| **generated query** | **cosine score** |
| If one gets Botox into the chin area to smooth out a pebbly chin, does the mentalis muscle eventually atrophy? | 0.9868196 |

Table 3: Generated queries are presented in conjunction with their respective passages, accompanied by cosine similarity scores. Notably, scores below 0.25 are indicative of query irrelevancy. While the filtration process effectively serves its purpose, it is important to acknowledge that passages yielding cosine scores close to 1.0 often are questions themselves.

| Dataset | Generated query |
|---|---|
| SciFact | How much of the p66 reverse transcriptase is processed in the cellular host?<br>sensitivity of HPV testing in primary screening for cervical abnormalities<br>ethanol extract of Allium fistulosum can be utilized as a drug for non-alcoholic fatty liver disease? |
| FEVER | What event was Musaeb Abdulrahman Balla in 2012?<br>Deokjeok Island is located in which county?<br>1993 northeastern conference baseball tournament |
| Climate-FEVER | What is the current temperature of Lost Creek, Pennsylvania?<br>Does cyclas balansae grow in a humid or wet climate?<br>What is the mean sea level of Roy Taylor? |
| SCIDOCS | What is the low cost dual polarized base station element for 4G LTE?<br>Gender and tenure diversity in the github team<br>Supply chain information systems strategy and its impact on firm performance. |
| ArguAna | Cannabis is not harmful. What are its benefits?<br>How would you pay for school uniform if you don't earn enough?<br>Custodial sentences can reduce recidivism and other associated problems. |
| NFCorpus | How do gut bacteria contribute to the obesity of humans?<br>is there an association between coffee or tea drinkers and a lower risk of depression?<br>Brachial arterial stiffness is lower in vegetarians than in omnivores. |
| DBPedia | 2014-15 a-1 league top teams<br>what event did charles lefrançois compete in?<br>what are the states in new york and pennsylvania? |
| FiQA | How do passive income income compare with active income?<br>How much can a W-2 employee deduct for their parking and commuter expense?<br>What are the options for filing an interest claim for a savings bank account? |
| Touché2020 | Does football have a harder game setting than rugby?<br>Animal testing is the only thing that makes medical research possible<br>Samsung Galaxy S5 is better than the iPhone 5S |
| HotpotQA | what is jaime silva gomez position in colombia football?<br>How does unbounded nondeterminism compare to indeterminacy?<br>How can I locate a book on boys working underground in coal mines in Canada? |

Table 4: Examples of generated queries with TOPiC. 3 examples are displayed per each BeIR dataset.