# Questions are Not All You Need for Brewing BeIR

Anonymous ACL submission

## Abstract

This paper studies the problem of information retrieval, to adapt to unseen tasks. Existing work generates synthetic queries from domain-specific documents to jointly train the retriever. However, the conventional query generator inadvertently or intentionally assumes the query as a *question*, thus failing to accommodate general search intents. A more lenient approach, which we refer to as *retrieval with intent*, incorporates task-specific elements into the query generation process, such as few-shot learning. In this paper, we explore novel strategies for task adaptation by guiding the LM to generate queries covering diverse **search intents**, using instructions and relevant demonstrations. We propose EGG, a query generator that better adapts to wide search intents expressed in the BeIR benchmark. Our method outperforms existing models on four tasks with underexplored intents, while utilizing 47 times smaller query generator compared to the previous state-of-the-art. Together, our work sheds light on how to integrate diverse search intents into the query generation process.

## 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 latent spaces within query and passage encoders (Gao and Callan, 2021; Ni et al., 2021; Santhanam et al., 2021). However, this requires collecting labeled data across numerous domains, while such labels are often unavailable.

In such scenarios, where only given the target corpus, existing work focuses on query generation to form a synthetic dataset (Cheriton, 2019; Ma et al., 2021). Representing approaches include GenQ (Thakur et al., 2021a), training a query generator using MSMARCO (Campos et al., 2016), an extensive question-answering dataset, and Promptagator-Zero (Dai et al., 2022), prompting the LLM to generate questions about the documents. With this convention, the former inadvertently generates queries in questions forms, being pretrained from MSMARCO queries in questions form, while the latter intentionally does so.

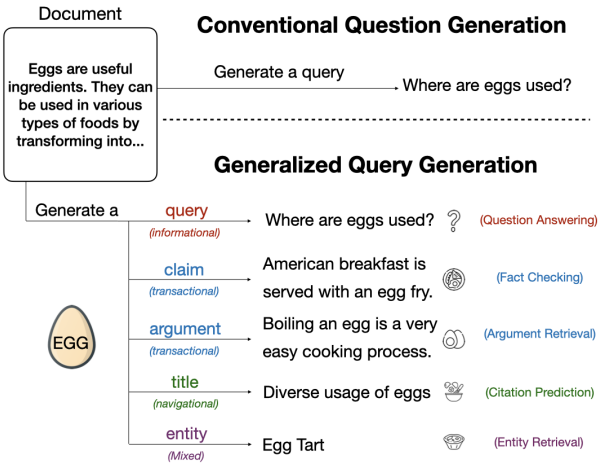


Figure 1: Overview of our method. Given a document, a conventional query generator generates a question. Our method reflects diverse search intents from various tasks to enhance the generalizability of the query generator. Fact checking can be viewed as *transaction*, where the retriever determines whether the given claim is supported or not, and argument retrieval is similarly so. Citation prediction, presenting a title as the query, represents a *navigational* intent for a specific document and entity retrieval exhibits a mixture of the intents.

To illustrate this point, Broder (2002) categorizes web search behaviours into three distinct intents: *Informational*, *Navigational*, and *Transactional*. Queries presented in question form predominantly align with the first category. While concentrating on the first category in the representative zero-shot retrieval BeIR (Thakur et al., 2021a) benchmark is a sensible strategy given the prevalence of datasets aligned with informational intent, we contend that recognizing diverse search intents is beneficial. This is illustrated with four

non-QA BeIR tasks resented in Figure 1, which better align with understudied intents.

To bridge these task gaps, recent work has attempted to incorporate task-specific elements, which we refer to as *retrieval with intent* (Asai et al., 2022; Hashemi et al., 2023). Promptagator-Few leverages FLAN 137B (Wei et al., 2021) as a few-shot query generator to better capture the latent intent in the query-document pairs. Nevertheless, in datasets like Climate-Fever (Diggelmann et al., 2020), where relevant pairs are not well-defined, in-context learning (Brown et al., 2020) suffers due to the poor quality of examples (Nori et al., 2023).

In this paper, we propose **EGG** (Qu**E**ry **G**enerator **G**eneralized), which generalizes the conventional query generation process by incorporating unique search intents. Our system comes in two model sizes. EGG-FLAN is for scenarios where the model is too small to support in-context learning, for which, we adapt instructions for diverse search intents. However, we find the LM generated queries lack sufficient diversity, thus we ensure the instruction to diversify those. When the model is large, we propose EGG-LLAMA, which improves the quality of demonstrations, query-document pairs, with high relevance. Specifically, we obtain *prototype* queries with instructions and utilize relevance ranking for retrieving demonstrations.

Our method not only enhances the baseline but also outperforms previous methods, while employing a 47 times smaller query generator compared to the state-of-the-art. EGG effectively covers undersupported search intents in BeIR, where existing methods often suffer. Our findings offer valuable insights into the query generation process, addressing the unique demands of search intents.

## 2 Related Work

In scenarios where only the target corpus is available (Izacard et al., 2021; Ni et al., 2021; Gao et al., 2022), existing works create synthetic labels by generating queries from documents. A common approach is to train a query generator on large QA datasets (Cheriton, 2019; Ma et al., 2021). Another line of study prompts LLMs to generate synthetic questions from documents and train a retriever (Sachan et al., 2022b), or a reranker (Sachan et al., 2022a; Bonifacio et al., 2022). Nevertheless, being evaluated on QA tasks, they intentionally generate questions to satisfy the search intent. More related to our setting, Promptagator leverages few-shot examples to capture the latent intents, while exemplars may not be efficient enough. Other work integrates search intents to retrieval without query generation, such as prepending user intents into the query (Asai et al., 2022) or introducing a scenario where only expert-written domain descriptions are available (Hashemi et al., 2023).

| Task | Intent | $e_q$ |
|---|---|---|
| Fact Checking | Transactional | Claim |
| Argument Retrieval | Transactional | Argument |
| Citation Prediction | Navigational | Title |
| Entity Retrieval | Mixed | Entity |

Table 1: Query description ($e_q$) of the tasks with underexplored search intents.

Concurrent to our work, UDAPDR (Saad-Falcon et al., 2023) generates a small number of queries using GPT-3 and iteratively generates a large number of queries with cheaper model to train rerankers. Compared to us, both iterative query generation and reranker training entail high costs. Moreover, our main focus is centered on integrating the unique search intent of each task, rather than producing good queries for the specific documents.

## 3 Methodology

### 3.1 Dataset

BeIR is a comprehensive benchmark for zero-shot retrieval, including 9 distinct tasks. We focus on tasks that involve underexplored intents. Specifically, we evaluate fact checking task Fever (Thorne et al., 2018), argument retrieval task Arguana (Wachsmuth et al., 2018), citation prediction task Scidocs (Cohan et al., 2020), and entity retrieval task DBPedia (Hasibi et al., 2017)[1].

### 3.2 Task Formulation

Given the corpus $D^c = \{d_i\}$ and the search intent $e_q$, the goal is to generate queries $\{q_i^*\}$ that correspond to $e_q$. Utilizing the synthetic pairs $\{d_i, q_i^*\}$, passage and query encoders are jointly trained to align their latent spaces. In conventional zero-shot approaches, $e_q$ is often considered as either the term 'query' (Dai et al., 2022; Wang et al., 2023; Saad-Falcon et al., 2023), or 'question' (Sachan et al., 2022a; Bonifacio et al., 2022), both resulting in the generation of questions as queries. We generalize this assumption by treating $e_q$ differently for each task. Unlike Promptagator, which relies on

[1]We select 1 dataset from each task with underexplored intent.

| | model size | | Fact | Argument | Citation | Entity | |
|---|---|---|---|---|---|---|---|
| | QG | retriever | Fever | Arguana | Scidocs | DBPedia | Avg. |
| *Unsupervised* | | | | | | | |
| BM25 | - | - | 75.3 | 31.5 | 15.8 | 31.3 | 38.5 |
| Contriever | - | 110M | 68.2 | 37.9 | 14.9 | 29.2 | 37.6 |
| ART | - | 220M | 72.4 | 32.2 | 14.4 | 36.3 | 38.8 |
| *Pretrain-based* | | | | | | | |
| GTR-XXL | - | 4.8B | 74.0 | 54.0 | 16.1 | 40.8 | 46.2 |
| TART | - | 1.5B | - | 51.5 | <u>18.7</u> | <u>46.8</u> | - |
| *Generation-based* | | | | | | | |
| GenQ | 220M* | 66M | 66.9 | 49.3 | 14.3 | 32.8 | 40.8 |
| GPL | 220M* | 66M | 75.9 | 55.7 | 16.9 | 38.4 | 46.7 |
| Promptagator-Zero | 137B | 110M | 76.2 | 53.8 | 16.3 | 36.4 | 45.7 |
| Promptagator-Few | 137B | 110M | 77.0 | 59.4 | 18.5 | 38.0 | 48.2 |
| *Baseline* | | | | | | | |
| DPR + FLAN-T5 | 3B | 66M | 61.2 | 55.9 | 16.7 | 32.3 | 41.5 |
| + Llama2 | 7B | 66M | 60.9 | 59.0 | 16.0 | 34.2 | 42.5 |
| *Proposed method (*EGG*)* | | | | | | | |
| DPR + EGG-FLAN | 3B | 66M | 69.5 | 60.1 | **18.6** | 33.6 | 45.5 |
| + EGG-LLAMA | 7B | 66M | 67.6 | **<u>61.2</u>** | 18.2 | 32.5 | 44.9 |
| GPL + EGG-FLAN | 3B | 66M | **<u>79.4</u>** | 58.7 | 16.9 | 40.0 | **<u>48.8</u>** |
| + EGG-LLAMA | 7B | 66M | 78.3 | 57.1 | 17.0 | **40.2** | 48.2 |

Table 2: Model performances across four BeIR tasks in nDCG@10. QG indicates the model size of query generator. (*) GenQ and GPL further finetune the generator on MSMARCO. Our methods outperform *Generation-based* models. Bold and underline indicates the best score among *Generation-based* models and all models respectively.

a task-specific template of few-shot examples, we interpret $e_q$ as the description of the query and incorporate it into the instruction. We adopt $e_q$ from the BeIR paper, as presented in Table 1.

### 3.3 Query Generation

**EGG-FLAN** FLAN-T5 (Chung et al., 2022) is known for its strong ability to follow instructions (Sun et al., 2023). We enhance the diversity of the generated queries through instructions. We find that when generating queries naively, the model tends to extract same phrases from the document regardless of $N$. To encourage diversity, we prompt the model to generate of its own words. For each $d_i$, we employ the following prompt to generate $N$ queries $\{q^*_{i_k}\}, k \in (1, N)$ per document: "`Write a {`$e_q$`} related to topic of the passage. Do not directly use wordings from the passage. {`$d_i$`}`". To enhance the quality, we apply a filtration mechanism, retaining only $\{q^*_{i_k}, d_i\}$ pairs that exhibit similarity above certain threshold.

**EGG-LLAMA** Large language models exhibit strong performance via in-context learning. Many studies have emphasized the significance of in-context examples relevant to the given document (Liu et al., 2021; Lee et al., 2023), where few-shot methods inevitably employ a fixed set, regardless of the document. To benefit from such relevance, we first generate *prototype* queries $\{q'_i\}$ and then demonstrate relevant examples $\{P_i\} = (d_i, q'_i)$. Initially, we generate one $q'_i$ per document with Llama2-chat (Touvron et al., 2023) using the following instruction: "`[INST] Read the passage and generate a {`$e_q$`}. [/INST] {`$d_i$`} {`$e_q$`} :`".

Subsequently, we perform in-context learning on Llama2 model with $\{P_i\}$ to obtain queries of high relevance and quality to the given document (Saad-Falcon et al., 2023). With respect to some similarity function $f(i, j)$, we retrieve $M$ relevant examples $P_{i_k} = (d_{i_k}, q'_{i_k}), k \in (1, M)$ for each document. Finally, we generate $\{q^*_{i_k}\}$ with the following prompt: "`Passage: {`$d_{i_1}$`} {`$e_q$`}: {`$q'_{i_1}$`} ... Passage: {`$d_{i_M}$`} {`$e_q$`}: {`$q'_{i_M}$`} Passage: {`$d_i$`} {`$e_q$`}:`" Since *prototype* queries are tailored to $e_q$, we can expect $\{q^*_{i_k}\}$ to also be aligned with $e_q$.

## 4 Experiments

We experiment with FLAN-T5-XL (3B) and Llama2 (7B) models, generating 8 queries per passage (qpp) and showcasing 4 examples during in-context learning. We train a DistilBERT TAS-B (Hofstätter et al., 2021) retriever with DPR

(Karpukhin et al., 2020) and GPL (Thakur et al., 2021b) training frameworks. For EGG-FLAN, we train an initial retriever and remove query-document pairs that exhibit cosine score less than 0.25, which was selected by experimenting on MS-MARCO (Figure 3). For EGG-LLAMA, we use the dot product between the SimCSE (Gao et al., 2021) embeddings of $d_i$ and $d_j$ as $f(i, j)$.

Given the different pipelines in existing works, we establish a baseline with DPR for a fair comparison. Following the conventional zero-shot assumption, we define $e_q$ as the term 'query' to generate questions. We evaluate with nDCG@10 metric, a standard measure for the BeIR benchmark. For further details, please refer to Appendix A.

## 5 Results

We present our results in Table 2.

### 5.1 EGG vs Baseline

Both EGG-FLAN and EGG-LLAMA improves the baseline by 4.0 and 2.4 nDCG scores. While Llama2 exhibits superior baseline performance, we observe that EGG-FLAN outperforms EGG-LLAMA. These results indicate that the queries generated by our method effectively cater to various search intents compared to questions.

### 5.2 EGG-FLAN vs EGG-LLAMA

We compare the results of EGG-FLAN and EGG-LLAMA here. Despite its smaller size and lower computational costs, EGG-FLAN with GPL demonstrates the highest average score. EGG with GPL shows a better average score than DPR, primarily due to the gains on Fever and DBPedia. Nevertheless, EGG with DPR exhibits marginal improvements in the other two tasks. Since our primary focus in this study is establishing an effective query generator, we leave incorporating more powerful retrievers or rerankers for future work.

### 5.3 Overall Performance

EGG-FLAN with GPL achieves the top rank among the *Generation-based* models, while EGG-LLAMA attains the second, tied with Promptagator-Few. While TART exhibits the highest performance among all models, our method performs better on Arguana and comparably on Scidocs. As *Generation-based* methods can adjust to unseen tasks using synthetic queries, our work explores leveraging unique search intents as a promising avenue to develop a more efficient query generator.

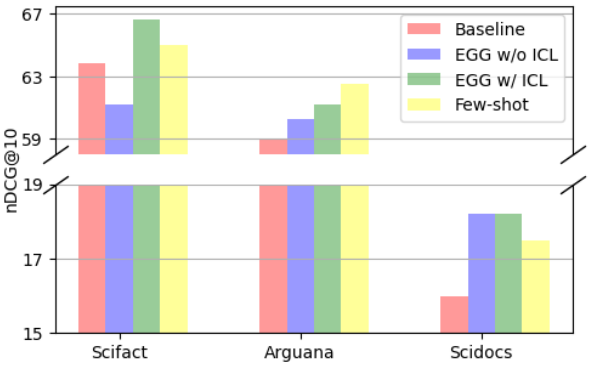


Figure 2: Retrieval performance of different query generation methods on three tasks. EGG w/o ICL indicates EGG-LLAMA model without performing additional in-context learning. For fact checking task, we experiment on SciFact (Wadden et al., 2020) dataset to observe if our findings can be generalized to other dataset.

### 5.4 Analysis

**Ablation Study** Since Llama2 model is proficient at few-shot learning, we conduct an ablation study across different query generation methods. We train a retriever with 8 *prototype* queries per document to examine the effect of in-context learning. Additionally, we leverage 4 random examples for few-shot learning. As illustrated in Figure 2, EGG-LLAMA exhibits superiority among others. Moreover, in-context learning enhances performance, particularly when the generated queries are longer (claim, argument vs title).

**Qualitative Analysis** Unlike baseline queries, queries generated with EGG are similar to few-shot and gold queries (Table 4). Meanwhile, as the original task of Arguana is to retrieve *counter argument*, we observe that EGG-generated query semantically contradicts with the gold query. Defining more specific search intent for each task may enable the construction of better synthetic queries.

## 6 Conclusion

In this work, we present EGG, a novel approach designed to overcome the shortcomings of previous query generation methods. We propose two designs for the query generator, distinguished by their model sizes, to improve the diversity and quality of synthetic queries while effectively capturing the search intent of the task. Our approach demonstrates superior performance across four tasks, suggesting a promising direction for query generation involving underexplored search intents.

## Limitations

Some tasks exhibit mixed search intents, such as DBPedia and NFCorpus (Boteva et al., 2016) datasets. While we have adopted the most representative description of the query, which is provided by the authors of BeIR, considering multiple candidates for $e_q$ has the potential to augment the performance. Moreover, our study focuses on the commonly affordable sizes of LMs. We reserve the exploration of larger variants of FLAN-T5 and Llama2 models to future investigations, seeking to discern their capacity to specialize in certain tasks. Lastly, utilizing a reranker instead of a retriever has demonstrated high performance (Dai et al., 2022; Bonifacio et al., 2022; Saad-Falcon et al., 2023), despite its expensive computations. The performance of our query generator has the potential to be further enhanced with the incorporation of the reranker.

## References


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*.

Luiz Henrique Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Unsupervised dataset generation for information retrieval. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

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*.

Andrei Z. Broder. 2002. A taxonomy of web search. *SIGIR Forum*, 36:3–10.

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.

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.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *ArXiv*, abs/2104.08821.

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*.

Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. 2017. Dbpedia-entity v2: A test collection for entity search. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

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*.

Yoonsang Lee, Pranav Atreya, Xi Ye, and Eunsol Choi. 2023. Crafting in-context examples according to lms' parametric knowledge. *ArXiv*, abs/2311.09579.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? In *Workshop on Knowledge Extraction and Integration for Deep Learning Architectures; Deep Learning Inside Out*.

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.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *ArXiv*, abs/2311.16452.

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.

Jon Saad-Falcon, O. Khattab, Keshav Santhanam, Radu Florian, Martin Franz, Salim Roukos, Avirup Sil, Md Arafat Sultan, and Christopher Potts. 2023. Udapdr: Unsupervised domain adaptation via llm prompting and distillation of rerankers. *ArXiv*, abs/2303.00807.

Devendra Singh Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen tau Yih, Joëlle Pineau, and Luke Zettlemoyer. 2022a. Improving passage retrieval with zero-shot question generation. In *Conference on Empirical Methods in Natural Language Processing*.

Devendra Singh Sachan, Mike Lewis, Dani Yogatama, Luke Zettlemoyer, Joëlle Pineau, and Manzil Zaheer. 2022b. Questions are all you need to train a dense passage retriever. *Transactions of the Association for Computational Linguistics*, 11:600–616.

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*.

Jiu Sun, Chantal Shaib, and Byron Wallace. 2023. Evaluating the zero-shot robustness of instruction-tuned language models. *ArXiv*, abs/2306.11270.

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, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Annual Meeting of the Association for Computational Linguistics*.

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.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. *ArXiv*, abs/2303.07678.

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.

# A Experimental Details

## A.1 Implementation

We utilize the publicly available FLAN-T5-XL and Llama2 checkpoints[2]. 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 100K 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. Generating queries with EGG-FLAN are conducted on a single RTX 3090 GPU and generating queries with EGG-LLAMA are conducted on 4 RTX A6000 GPUs. Query generation with EGG-FLAN took 15 hours in total and EGG-LLAMA took 75 hours in total. Training with DPR took maximum 1-2 hours per each dataset. We did not modify any training pipeline of GPL[3].

## A.2 Cosine Filtering

In this section, we describe the filtration mechanism for EGG-FLAN. We train an initial retriever with the synthetic dataset for 1 epoch. Subsequently, cosine product for each document and synthetic query is measured using the trained retriever. We remove all pairs that exhibit cosine similarity scores below 0.25. The particular threshold value is selected through experiments on MSMARCO, as shown in Figure 3. We further present examples and the corresponding scores for MSMARCO in Table 3.

## A.3 Dense Retrieval

## A.4 Evaluation

Among *unsupervised* models, we benchmark against BM25, Contriever (Izacard et al., 2021), and ART(Sachan et al., 2022b). As *Generation-based* models, GenQ and GPL employ MSMARCO-trained T5 (Raffel et al., 2019) model as the query generator and finetune TAS-B model. Promptagator leverages FLAN 137B as the generator and GTR-base as the retriever. While *Pretrain-based* models not being our main scope of study, GTR (Ni et al., 2021) and TART (Asai et al., 2022) train the retriever on multiple tasks without accessing the target corpus. The performance is boosted by scaling up the retriever size (GTR) or utilizing a cross-encoder as the retriever (TART).

[2]https://huggingface.co/google/flan-t5-xl, https://huggingface.co/meta-llama/Llama-2-7b-hf, https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

[3]https://github.com/UKPLab/gpl

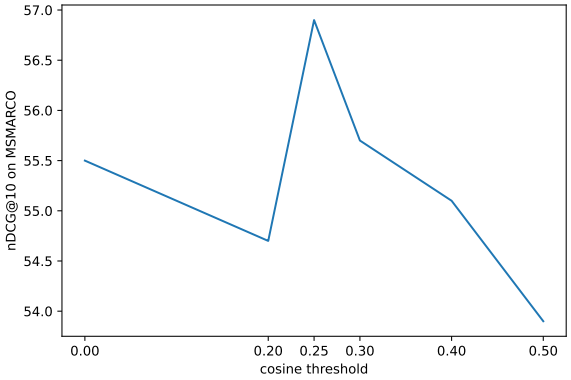


Figure 3: Performance on MSMARCO by differing cosine thresholds. Generated queries of cosine score lower than each value of x-axis is removed.

# B Qualitative Analysis

In this section, we compare the generated queries to the gold query for the experiment in Section 5.4 (Table 4). Overall, we observe that baseline queries exhibit a question form, which is far apart from gold queries. On the other hand, EGG and few-shot queries demonstrate a similar form to the gold queries. The results indicate the efficacy of our method as the few-shot learning, which has already demonstrated its strong performance on various tasks.

Furthermore, we provide examples of 3 generated queries for each dataset. Table 5 and 6 presents queries generated with EGG-FLANand EGG-LLAMA, respectively.

| 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 | Baseline | What is the relationship between the biomarkers of ovarian reserve and the risk of infertility among older women of reproductive age? |
| | EGG w/o ICL | The use of biomarkers of ovarian reserve, specifically antim00fcllerian hormone (AMH), follicle-stimulating hormone (FSH), and inhibin B, is not a reliable indicator of reproductive potential among women of late reproductive age, |
| | **EGG w/ ICL** | A low level of ovarian hormones is associated with an increased likelihood of infertility in women of late reproductive age. |
| | Few-shot | AMH levels are associated with the likelihood of successful fertility treatment outcomes. |
| | **Gold** | A diminished ovarian reserve does not solely indicate infertility in an a priori non-infertile population. |
| Arguana | Baseline | What are the ethical and environmental impacts of meat production and consumption, and how do they compare to plant-based food systems? |
| | EGG w/o ICL | "Do vegetarians really help the environment?", Simon Farrell offers an opposing view, arguing that vegetarianism may not be the most effective way to reduce the environmental impact of food production. According to Farrell, livestock farming creates unique environments, such as chalk down land and mountain |
| | **EGG w/ ICL** | Growing crops and vegetation can also be bad for the environment due to the use of fertilisers and pesticides and the destruction of forests and wildlife habitats, as well as the inefficiency of agricultural land use in some countries. |
| | Few-shot | The meat industry is not green While many vegetarians and vegans may think that their diet is environmentally friendly, the reality is that the meat industry is not green. Meat production is a major contributor to greenhouse gas emissions, deforestation, and water pollution. |
| | **Gold** | Being vegetarian helps the environment Becoming a vegetarian is an environmentally friendly thing to do. Modern farming is one of the main sources of pollution in our rivers. Beef farming is one of the main causes of deforestation, and as long as people continue to buy fast food in their billions, there will be a financial incentive to continue cutting down trees to make room for cattle. (...) |
| Scidocs | Baseline | What are the key differences between RMSProp and ESGD adaptive learning rate schemes for non-convex optimization? |
| | EGG w/o ICL | Equilibration Preconditioner and Adaptive Learning Rate Schemes for Non-Convex Optimization |
| | **EGG w/ ICL** | Equilibration-Based Learning Rates for Deep Neural Networks with Non-Convex Losses |
| | Few-shot | Escape from Saddle Points: A Survey of Non-Convex Optimization Techniques for Deep Learning |
| | **Gold** | Train longer, generalize better: closing the generalization gap in large batch training of neural networks |

Table 4: Examples of generated queries from the ablation study in Section 5.4. Baseline indicated the query generator that generates questions. EGG w/o ICL indicates EGG-LLAMA without in-context learning stage. We observe that EGG demonstrates high similarity to gold queries, with respect to $e_q$.

| Dataset | Generated query |
|---|---|
| Fever | Monochamus adamitus is a species of beetle in the Cerambycidae family.<br>Der Klassiker is the name given in German to the match between two German football clubs.<br>Bootstrapping populations for parametric inference. |
| Arguana | Animals are sentient beings who can feel pleasure and pain. Animal suffering is just as serious as human suffering. Therefore it is immoral to kill animals for food when we do not need to do so.<br>Sport and politics are separate and should be kept separate<br>The Heathrow Airport has been at capacity since it was built and will continue to be. |
| Scidocs | WhatsApp Usage Patterns and Prediction Models<br>Random Walk with Restart on Large Graphs Using Block Elimination<br>Context Suggestion for User-Oriented Recommender Systems |
| DBPedia | Jindo Island<br>Game of Thrones (season 5)<br>2002 IIHF World Junior Ice Hockey Championships |

Table 5: Examples of generated queries with EGG-FLAN. 3 examples are displayed per each BeIR dataset.

| Dataset | Generated query |
|---|---|
| Fever | Monochamus adamitus is threatened by habitat loss and fragmentation, which is caused by logging and agricultural activities.<br>The matches between Bayern Munich and Borussia Dortmund are considered to be some of the biggest and most exciting in the German football leagues.<br>A bootstrap is a technique for approximating an unknown population distribution from a known sample drawn from it. |
| Arguana | Killing animals for food is unjustified and unnecessary, and can be replaced with plant-based or lab-grown alternatives that do not require the killing of animals.<br>The Euro 2012 football tournament should not be used for political posturing and grandstanding.<br>Heathrow airport must expand in order to maintain its competitiveness and avoid falling behind other European airports. |
| Scidocs | A Study of WhatsApp Messaging and Behavior: Predictive Models and User Characteristics<br>Fast and Accurate Random Walks with Restarts on Large Graphs Using Block Elimination<br>Context Suggestion: User-Oriented Context Recommendation in Recommender Systems |
| DBPedia | Battle of Myeongnyang<br>Game of Thrones Season 5<br>2002 Men's World Ice Hockey Championships |

Table 6: Examples of generated queries with EGG-LLAMA. 3 examples are displayed per each BeIR dataset.