Disentangling Questions from Query Generation for Task-Adaptive Retrieval

Yoonsang Lee Minsoo Kim Seung-won Hwang*
Seoul National University
{lysianthus, minsoo9574, seungwonh}@snu.ac.kr

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 assumes the query as a question, thus failing to accommodate general search intents. A more lenient approach incorporates task-adaptive elements, such as few-shot learning with an 137B LLM. In this paper, we challenge a trend equating query and question, and instead conceptualize query generation task as a "compilation" of high-level intent into task-adaptive query. Specifically, we propose EGG, a query generator that better adapts to wide search intents expressed in the BeIR benchmark. Our method outperforms baselines and existing models on four tasks with underexplored intents, while utilizing a query generator 47 times smaller than the previous state-of-the-art. Our findings reveal that instructing the LM with explicit search intent is a key aspect of modeling an effective query generator.[2]

1 Introduction

Information retrieval has significantly facilitated the process of locating relevant documents in response to user requests. With the advent of dense retrieval [23], prior works have concentrated on the supervised alignment of latent spaces within query and passage encoders [14, 29, 35]. However, this requires collecting labeled data across numerous domains, which are often unavailable. In such scenarios, where only given the target corpus, existing work focuses on zero-shot query generation to form a synthetic dataset [10, 27]. Representing approaches include GenQ [37], training a query generator using MSMARCO [8], an extensive question-answering dataset, and Promptagator-Zero [13], prompting the LLM to generate questions about the documents. While the former inadvertently generates queries in questions forms, being pretrained from MSMARCO queries in questions form, the latter intentionally does so.

We challenge the prevailing trend of equating queries with questions in training. Though question form may align with information intent, web search intents are categorized to include more diverse intents [6] such as *Informational*, *Navigational*, and *Transactional*. While concentrating on the first category has paid off, given the prevalence of datasets aligned with informational intent, generalization is critical, by four non-QA datasets in BeIR [37] tasks in Figure 1.

We thus interpret query generation task as a "compilation" of high-level intent into task-adaptive query, to reflect diverse search intents [3, 2, 17]. This task has been mostly delegated to in-context learning: Promptagator-Few leverages FLAN 137B [43] as a few-shot query generator to better capture the latent intent in query-document pairs. However, challenges may arise in in-context

*Corresponding author.
[2]Our code is available at https://github.com/lilys012/metaprompt-QG.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

learning [7] due to small model size [44], limited context length [25], or poor quality of examples [30].

To overcome the reliance on expensive few-shot examples and enable the use of small LMs as query generators, we propose **EGG** (**E**fficient **G**eneralized **G**enerator), which leverages *meta-prompt*[3] to incorporate 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. In cases where the model size is just enough to support in-context learning (EGG-LLAMA), we first generate *prototype* queries with meta-prompt as candidates for in-context query-document pairs. Another key distinction is, we utilize retriever feedback, specifically, relevance ranking. EGG not only surpasses both zero-shot and few-shot baselines but also outperforms previous state-of-the-art methods. Our method effectively covers under-supported intents which existing methods often struggle with.

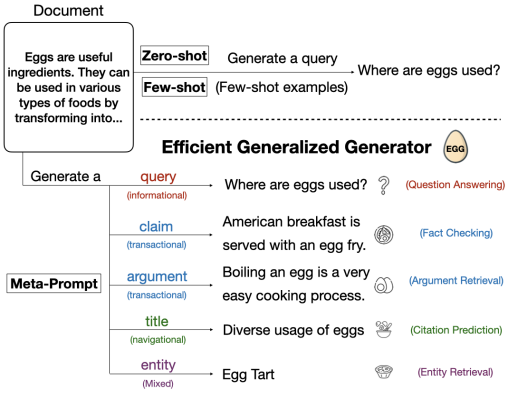

Figure 1: Overview of our method. Given a document, conventional zero-shot query generator generates questions, while few-shot query generator performs in-context learning with few-shot examples. In contrast, our method reflects diverse search intents utilizing meta-prompts to enhance the generalizability 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.

2 Methodology

2.1 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 well to e_q. Equipped with the synthetic pairs $\{d_i, q_i^*\}$, passage[4] and query encoders are jointly trained to align their latent embedding spaces. During inference, documents and test queries are each embedded with the trained encoders, and the top-k documents with the highest scores are retrieved.

In conventional zero-shot methods, e_q is often considered as either the term 'query' [13, 42, 32], or 'question' [33, 3], interchangeably. We challenge this assumption by defining e_q as query adaptive to each task, by interpreting e_q as task-specific attribute incorporated it into the meta-prompt.

2.2 Dataset

BeIR is a comprehensive benchmark for zero-shot retrieval, including 9 distinct tasks. We focus on tasks that involve underexplored intents and select one dataset from each task. Specifically, we evaluate fact checking task Fever [38], argument retrieval task Arguana [40], citation prediction task Scidocs [12], and entity retrieval task DBPedia [18]. We adopt e_q from the BeIR paper, as presented in Table 1.

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.

2.3 Query Generation

EGG-FLAN FLAN-T5 [11] is known for its strong ability to follow instructions [36]. For each d_i, we employ the following meta-prompt to generate N queries $\{q_{i_k}^*\}, k \in (1, N)$ per

[3]Following Nayak et al. [28], we define *meta-prompt* as the prompt that accepts a document and a task attribute as input to generate a task-adaptive query.

[4]We use both *document* and *passage* interchangeably.

		Fever	Arguana	Scidocs	Dbpedia	Avg.
DPR	FLAN (zero)	54.4	53.9	14.8	30.4	38.4
	FLAN (few)	38.1	22.9	16.7	31.5	27.3
	EGG-FLAN	**69.5**	**60.1**	**18.6**	**33.6**	**45.5**
	Llama2 (zero)	60.9	59.0	16.0	**34.2**	42.5
	Llama2 (few)	67.2	**62.0**	17.5	30.2	44.2
	EGG-LLAMA	**67.6**	61.2	**18.2**	32.5	**44.9**
GPL	FLAN (zero)	73.2	55.8	14.9	39.6	45.9
	FLAN (few)	77.0	3.5	16.3	37.7	33.6
	EGG-FLAN	**79.4**	**58.7**	**16.9**	**40.0**	**48.8**
	Llama2 (zero)	73.0	56.4	15.0	**40.2**	46.2
	Llama2 (few)	**79.8**	55.8	**17.0**	39.1	47.9
	EGG-LLAMA	78.3	**57.1**	**17.0**	**40.2**	**48.2**

Table 2: Results of EGG, zero-shot, and few-shot baselines with two training methods, DPR and GPL. EGG consistently outperforms the baselines. We bold the highest score for each experiment. Fev, Arg, Sci, Dbp stand for Fever, Arguana, Scidocs, and DBPedia.

document: "`Write a {`e_q`} related to topic of the passage. Do not directly use wordings from the passage.` $\{d_i\}$". Generated queries can reflect the search intent e_q through the instruction for better alignment of the retrievers. Our pilot study showed that when FLAN-T5 is prompted without specific instruction to write queries in its own words, it tends to extract the same sentence from the document as N queries, leading to worse performance. Therefore, we have added such instruction to encourage the model to attempt paraphrasing and explore various parts of the document.

Our meta-prompt incurs the same computational cost as the zero-shot prompt, which has 'query' as the task attribute. Compared to the few-shot method, meta-prompt is significantly shorter than the concatenation of few-shot examples, and even shorter than a single document, highlighting the efficiency aspect.

EGG-LLAMA LLMs can exhibit strong performance via in-context learning, especially when in-context examples are relevant to the given context [26, 24]. However, few-shot methods inevitably employ a fixed set of in-context examples, regardless of the documents. In contrast, we enable in-context learning with *prototype* queries generated by meta-prompts to benefit from relevant documents. Concretely, we first generate prototype queries $\{q'_i\}$ that align closely with the search intent of the target task, and then perform in-context learning using these relevant examples $\{P_i\} = (d_i, q'_i)$.

Initially, we generate one q'_i per document with Llama2 model [39] using the following meta-prompt: "`[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 [32]. 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 template: "`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.

3 Experiments

3.1 Experimental Setup

We experiment with FLAN-T5-XL (3B) and Llama2 (7B) models, generating 8 queries per passage and showcasing 4 examples during in-context learning. For EGG-LLAMA, we use the dot product between the SimCSE [16] embeddings of d_i and d_j as $f(i, j)$. We train a DistilBERT TAS-B [20] retriever for each task.

Among unsupervised models, we benchmark against BM25, Contriever [22], and ART [34]. For Generation-based models, we report GenQ, GPL, and two versions of Promptagator. While Pretrain-based models are not our main focus, we also include two representatives GTR [29] and TART

	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	18.7	46.8	-
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
DPR + EGG-FLAN	3B	66M	69.5	60.1	**18.6**	33.6	45.5
+ EGG-LLAMA	7B	66M	67.6	**61.2**	18.2	32.5	44.9
GPL + EGG-FLAN	3B	66M	**79.4**	58.7	16.9	40.0	**48.8**
+ EGG-LLAMA	7B	66M	78.3	57.1	17.0	**40.2**	48.2

Table 3: Model performances across four BeIR tasks in nDCG@10. QG and retriever indicates the model size of query generator and retriever. DPR+EGG-FLAN represents the retriever trained on EGG-FLAN-generated queries with DPR. Bold and underline indicate the best score among Generation-based models and all models. (*) GenQ and GPL further finetune the generator on MSMARCO. (**) Fever is included in the train corpus of TART.

[2]. These methods train a large retriever on massive pretrain corpus with multiple tasks but do not finetune on the target corpus. We evaluate with nDCG@10 metric, a standard measure for the BeIR benchmark. For further details, please refer to Appendix A.

3.2 Baselines

Considering different pipelines in existing works, we establish zero-shot and few-shot baselines for a controlled study. Following the conventional zero-shot assumption, we define e_q as the term 'query' to generate questions. For few-shot baseline, we employ 4-shot examples for Llama2 and 1-shot example for FLAN-T5, given its limited context size. We follow the same template used in in-context learning for EGG-LLAMA.

4 Results

4.1 Main Results

We compare the baselines and our method in Table 2. Both EGG-FLAN and EGG-LLAMA exhibit the highest performance among the baselines, underscoring the benefits of incorporating search intents through meta-prompts. Moreover, EGG-FLAN demonstrates a significant gain over its few-shot baseline, as small LMs may struggle to handle few-shot examples. These results indicate that the queries generated by our method effectively cater to various search intents.

While Llama2 exhibits superior baseline performance than FLAN-T5, EGG-FLAN outperforms EGG-LLAMA, despite its smaller size and lower costs. Additionally, GPL training method shows a better average score than DPR, primarily due to the gains in Fever and DBPedia. Nevertheless, EGG with DPR exhibits marginal improvements in the other two tasks.

4.2 Overall Performance

Table 3 describes the overall performance. EGG-FLAN with GPL achieves the top rank overall, while EGG-LLAMA with GPL attains the second, tied with Promptagator-Few. Compared to Pretrain-based models, our method performs better on Arguana and Fever, and comparably on Scidocs, despite much smaller retriever size and fewer pretrain tasks. As Generation-based methods can adjust to

unseen tasks using synthetic queries, our work suggests a promising avenue for developing a more versatile query generator.

4.3 Analysis

Ablation Study We conduct an ablation study on the effect of in-context learning in EGG-LLAMA. We train the retriever with 8 *prototype* queries without performing in-context learning. As illustrated in Table 4, in-context learning enhances performance, especially when the generated queries are longer and in a form of sentences (claim, argument vs. title, entity).

Fever	Arguana	Scidocs	DBPedia
61.9 (-5.7)	60.3 (-0.9)	18.0 (-0.2)	31.9 (-0.6)

Table 4: Results of retrievers trained with DPR on 8 prototype queries. Performance drops across all datasets compared to DPR+EGG-LLAMA. We put the difference in parenthesis and colour it red.

Qualitative Analysis As shown in Table 5 in the appendix, we observe that zero-shot queries are in a question form, which is semantically far apart from gold queries. On the other hand, EGG-generated (EGG-LLAMA) and few-shot queries demonstrate similar form and content to the gold queries, illustrating the efficacy of our method on par with few-shot learning. Meanwhile, since Arguana retrieves counter arguments, we find EGG-generated queries semantically contradict the gold query. Exploring more specific intents may enable constructing of better synthetic queries.

5 Related Work

In scenarios where only the target corpus is available [22, 29, 15], existing works create synthetic labels by generating queries from documents. A common approach is to train a query generator on large QA datasets [10, 27].

Recent work prompts LLMs to generate synthetic queries from documents. There has been active research on training neural rerankers with synthetic queries [33, 9, 21]. InPars [3] performs few-shot learning, while InPars-Light [5] extends InPars by employing cost efficient recipes. UDAPDR [32] generates a small number of queries using GPT-3 and iteratively generates a large number of queries with cheaper model to train rerankers. Almeida and Matos [1] leverages small LMs to generate questions starting with common initiators, such as What, How, and When. Meanwhile, training task-adaptive retriever is underexplored, only being studied by Promptagator [13], which leverages few-shot examples to capture the latent intents, albeit at much higher computational costs.

Another line of study provides specific instructions along with the queries and builds a general-purpose retriever [2, 31, 45]. This involves large sizes of LLMs during inference, while our method establishes small, task-specific retrievers.

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 EGG, distinguished by their model sizes, to improve the diversity and quality of synthetic queries while effectively capturing the search intent of the task using meta-prompts. 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 [4] datasets. While we have adopted the most representative description of the query, provided by the authors of BeIR, considering multiple candidates for e_q may augment the performance. Moreover, our study focuses on the commonly affordable sizes of LMs. We reserve the exploration of other 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 [13, 3, 32], despite its expensive computations. The performance of our method could be further enhanced with the incorporation of the reranker.

References

[1] Tiago Almeida and Sérgio Matos. Exploring efficient zero-shot synthetic dataset generation for information retrieval. In *Findings*, 2024.

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

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

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

[5] Leonid Boytsov, Preksha Patel, Vivek Sourabh, Riddhi Nisar, Sayan Kundu, R. Ramanathan, and Eric Nyberg. Inpars-light: Cost-effective unsupervised training of efficient rankers. *ArXiv*, abs/2301.02998, 2023.

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

[7] 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. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

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

[9] Ramraj Chandradevan, Kaustubh D. Dhole, and Eugene Agichtein. Duqgen: Effective unsupervised domain adaptation of neural rankers by diversifying synthetic query generation. 2024.

[10] David R. Cheriton. From doc2query to doctttttquery. 2019.

[11] 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. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416, 2022.

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

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

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

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

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

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

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

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

[20] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy J. Lin, and Allan Hanbury. 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*, 2021.

[21] Chao-Wei Huang and Yun-Nung Chen. Instupr : Instruction-based unsupervised passage reranking with large language models. *ArXiv*, abs/2403.16435, 2024.

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

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

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

[25] Mukai Li, Shansan Gong, Jiangtao Feng, Yiheng Xu, Jinchao Zhang, Zhiyong Wu, and Lingpeng Kong. In-context learning with many demonstration examples. *ArXiv*, abs/2302.04931, 2023.

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

[27] Ji Ma, Ivan Korotkov, Yinfei Yang, Keith B. Hall, and Ryan T. McDonald. Zero-shot neural passage retrieval via domain-targeted synthetic question generation. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2021.

[28] Nihal V. Nayak, Yiyang Nan, Avi Trost, and Stephen H. Bach. Learning to generate instruction tuning datasets for zero-shot task adaptation. *ArXiv*, abs/2402.18334, 2024.

[29] 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. Large dual encoders are generalizable retrievers. *ArXiv*, abs/2112.07899, 2021.

[30] 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. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *ArXiv*, abs/2311.16452, 2023.

[31] Hanseok Oh, Hyunji Lee, Seonghyeon Ye, Haebin Shin, Hansol Jang, Changwook Jun, and Minjoon Seo. Instructir: A benchmark for instruction following of information retrieval models. *ArXiv*, abs/2402.14334, 2024.

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

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

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

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

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

[37] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 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)*, 2021.

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

[39] 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. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023.

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

[41] Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. Gpl: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. *ArXiv*, abs/2112.07577, 2021.

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

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

[44] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed Huai hsin Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *ArXiv*, abs/2206.07682, 2022.

[45] Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. Followir: Evaluating and teaching information retrieval models to follow instructions. *ArXiv*, abs/2403.15246, 2024.

A Experimental Details

A.1 Implementation

We utilize the publicly available FLAN-T5-XL and Llama2 checkpoints[5]. When generating prototype queries, since we provide instructions, we leverage the chat variant of Llama2 as it is known to be finetuned on instructions. When performing in-context learning, we employ the base model. 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.[6]

A.2 Retriever Training

As we have obtained a task-specific synthetic dataset, we now train a retriever on this dataset. We employ two training frameworks: DPR [23] and GPL [41]. DPR maximizes the likelihood of the product between query and positive document embeddings through in-batch negatives, while GPL soft-labels the score of query and positive document using cross-encoder and employs MarginMSE loss [19].

B Query Examples

We provide examples of 3 generated queries for each dataset. Table 6 and 7 presents queries generated with EGG-FLANand EGG-LLAMA, respectively.

[5]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

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

Dataset	Generated query	
Fever	Zero-shot	Who is the author of the novel "Don Quixote"?
	Few-shot	Sancho Panza's novel Don Quixote written by Don Miguel de Cervantes Saavedra.
	EGG w/o ICL	"Sancho Panza's character provides a commentary on the societal issues of the time, particularly those related to class and social hierarchy.
	EGG	Sancho Panza is a humorous and insightful character in the novel Don Quixote who provides commentary on the events of the story and offers a unique perspective on the world.
	Gold	Sancho Panza is a fictional character in a novel written by a Spanish writer born in the 17th century.
Arguana	Zero-shot	What are the ethical and environmental impacts of meat production and consumption, and how do they compare to plant-based food systems?
	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.
	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	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.
	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	Zero-shot	What are the key differences between RMSProp and ESGD adaptive learning rate schemes for non-convex optimization?
	Few-shot	Escape from Saddle Points: A Survey of Non-Convex Optimization Techniques for Deep Learning
	EGG w/o ICL	Equilibration Preconditioner and Adaptive Learning Rate Schemes for Non-Convex Optimization
	EGG	Equilibration-Based Learning Rates for Deep Neural Networks with Non-Convex Losses
	Gold	Train longer, generalize better: closing the generalization gap in large batch training of neural networks
DBPedia	Zero-shot	What is the population density of Steele Township based on the 2010 census?
	Few-shot	township Rowan county North Carolina United States
	EGG w/o ICL	Steele Township
	EGG	Steele Township, Rowan County, North Carolina
	Gold	rowan university

Table 5: Examples of generated queries with different methods. EGG w/o ICL indicates EGG-LLAMA without in-context learning stage, while EGG indicates the full EGG-LLAMA. We truncate too long queries due to space limit. We observe that the queries generated with our method demonstrate 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. Der Klassiker is the name given in German to the match between two German football clubs. 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. Sport and politics are separate and should be kept separate The Heathrow Airport has been at capacity since it was built and will continue to be.
Scidocs	WhatsApp Usage Patterns and Prediction Models Random Walk with Restart on Large Graphs Using Block Elimination Context Suggestion for User-Oriented Recommender Systems
DBPedia	Jindo Island Game of Thrones (season 5) 2002 IIHF World Junior Ice Hockey Championships

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

Dataset	Generated query
Fever	Monochamus adamitus is threatened by habitat loss and fragmentation, which is caused by logging and agricultural activities. The matches between Bayern Munich and Borussia Dortmund are considered to be some of the biggest and most exciting in the German football leagues. 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. The Euro 2012 football tournament should not be used for political posturing and grandstanding. 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 Fast and Accurate Random Walks with Restarts on Large Graphs Using Block Elimination Context Suggestion: User-Oriented Context Recommendation in Recommender Systems
DBPedia	Battle of Myeongnyang Game of Thrones Season 5 2002 Men's World Ice Hockey Championships

Table 7: Examples of generated queries with EGG-LLAMA. 3 examples are displayed per each dataset.