When do Generative Query and Document Expansions Fail? A Comprehensive Study Across Methods, Retrievers, and Datasets

Orion Weller[*ι] **Kyle Lo**[α] **David Wadden**[α] **Dawn Lawrie**[ι]
Benjamin Van Durme[ι] **Arman Cohan**[γα] **Luca Soldaini**[α]

[ι]Johns Hopkins University [α]Allen Institute for AI [γ]Yale University

oweller@cs.jhu.edu {kylel, lucas}@allenai.org

Abstract

Using large language models (LMs) for query or document expansion can improve generalization in information retrieval. However, it is unknown whether these techniques are universally beneficial or only effective in specific settings, such as for particular retrieval models, dataset domains, or query types. To answer this, we conduct the first comprehensive analysis of LM-based expansion. We find that there exists a strong negative correlation between retriever performance and gains from expansion: expansion improves scores for weaker models, but generally harms stronger models. We show this trend holds across a set of eleven expansion techniques, twelve datasets with diverse distribution shifts, and twenty-four retrieval models. Through qualitative error analysis, we hypothesize that although expansions provide extra information (potentially improving recall), they add additional noise that makes it difficult to discern between the top relevant documents (thus introducing false positives). Our results suggest the following recipe: use expansions for weaker models or when the target dataset significantly differs from training corpus in format; otherwise, avoid expansions to keep the relevance signal clear.[1]

Figure 1: LM-based query and document expansion methods typically improve performance when used with weaker models, but not for stronger models. More accurate models generally lose relevance signal when expansions are provided. Each point is a value in Table 1.

1 Introduction

Neural information retrieval (IR) systems routinely achieve state-of-the-art performance on tasks where labeled data is abundant (Karpukhin et al., 2020; Yates et al., 2021). When limited or no data is available, neural models fine-tuned on data-rich domains are used in zero-shot manner (Thakur et al., 2021; Rosa et al., 2022b). However, shifts in distribution of queries and documents can negatively impact their performance (Lupart et al., 2023).

To mitigate this effect, language models (LMs) can be used to *expand* queries or documents from unseen domains (Dai et al., 2022; Gao et al., 2022; Jagerman et al., 2023; Jeronymo et al., 2023; Wang et al., 2023a). These techniques input queries and/or documents into an LM to generate additional content, which is combined with original text to facilitate relevance matching. For example, Doc2Query (Nogueira et al., 2019c) uses an LM to generate likely queries for documents in the collection. Meanwhile, HyDE (Gao et al., 2022) uses an LM to generate a fictitious relevant document for a user query. As LMs are often trained on more domains than typical rankers, LM-based expansion leverages this encoded knowledge to bridge out-of-distribution gaps.

IR researchers have long proposed methods to expand queries and documents (Rocchio Jr, 1971; Lavrenko and Croft, 2001; Abdul-Jaleel et al., 2004). However, we note that LM-based expansions are qualitatively different from traditional

[1]Code and data are available at https://github.com/orionw/LM-expansions
[*] Work performed during internship at AI2.

Findings of the Association for Computational Linguistics: EACL 2024, pages 1987–2003
March 17-22, 2024

expansion techniques. While the latter are largely non-parametric, using thesauri or relevance signals from the collection,[2] LM-based expansions can leverage knowledge encoded in their model weights. Finally, while many comparative analyses of statistical expansion techniques exist (Hust et al., 2003; Bhogal et al., 2007; Carpineto and Romano, 2012), no equivalent work has been conducted for LM-based approaches.

Many works have proposed specific LM-based expansions, but these approaches are generally tested only a small subset of retrieval methods (small bi-encoder models or BM25) or only work on specific domains (Gao et al., 2022; Wang et al., 2023a; Zhu et al., 2023). We thus seek to answer the following:

RQ1: How do different models impact query and document expansion (§3)? Across all types of IR models and architectures, performance is negatively correlated with gains from expansion: after a certain score threshold these expansions generally hurt performance (as they blur the relevance signal from the original documents).

RQ2: How do different distribution shifts impact these results (§4)? Our main results hold for all types of shift – better models are harmed by expansion – except for long query shift, where expansions generally help most-to-all models.

RQ3: Why do expansions hurt stronger IR models (§5)? We find that query and document expansions introduce new terms, potentially weakening the relevance signal of the original text.

Overall, this work aims at answering the following question: **when should one use LM-based expansions?** Through our investigation, we provide evidence to help practitioners answer this question. Our results run counter to the common intuition that query and document expansion are helpful techniques in all cases; instead, they show that LM expansions generally **benefit weaker rankers**, but **hurt more accurate rankers**. Further, analysis over twelve datasets shows that whether a given model benefits from expansion varies depending on task; datasets with pronounced distribution shifts (*e.g.*, very long queries) are more likely to benefit.

[2]For example, pseudo relevance feedback (PRF) uses top-k retrieved documents to expand queries. Thus, PRF relies on the quality of the initial retrieved set; generally, the better the retrieval, the better the expansion. We note that this is not necessarily the case for LM-based expansions/PRF: parametric knowledge encoded in model weights affect terms selected for expansion (in contrast to classic PRF that typically selects new terms from the top relevant documents from the collection).

2 Experimental Settings

We provide an overview of document and query expansion methods used in the reminder of the manuscript, and describe our experimental setup.

We choose expansion techniques according to two criteria: (*i*) their overall performance, as claimed in papers introducing them, and (*ii*) whether they can used with any retrieval model. While there exists more specific techniques for particular architectures, such as ColBERT-PRF (Wang et al., 2023c,b), we use text-based expansions from LMs to ensure generalizability of our findings.

We generate expansions using `gpt-3.5-turbo`[3] as it is inexpensive and shows strong performance in previous work (Wang et al., 2023a; Jagerman et al., 2023). Since using LMs to generate expansions for large collections would be prohibitive, we restrict our expansions to the reranking setting, e.g. the top 100 documents per query found from BM25 following Asai et al. (2022).[4] Following established practices, we use these expansions for zero-shot out-of-domain retrieval. Although it is possible that training with expansions may further increase their effectiveness, this limits their generalizability since it requires re-training retrieval models for each expansion technique and LM.

2.1 Query Expansion

We use three types of query expansion, selecting the best methods from previous work.

HyDE from Gao et al. (2022) provides task-specific instructions for the LM to generate a document that would answer that question. We use prompts from their work when available.

Chain of Thought from Jagerman et al. (2023) was inspired by Wei et al. (2022); it prompts the model to reason before giving the answer. The step-by-step reasoning is then used to expand the

[3]We use version `gpt-3.5-turbo-0613`. To show that our results generalize beyond this specific language model, we include results using other open/API LMs (`gpt-4-0613`, Claude V2, Llama2 70b Chat) in Appendix A that show the same conclusion. Prompts and example input/output can be found in Appendix D and E. We also explore the placement of these augmentations (should we prepend/append/replace the original query and documents?) in Appendix B and show that this also makes little difference.

[4]As of September 2023, even just a single document expansion method using `gpt-3.5-turbo` on the DL Track 2019 collection would cost thousands of dollars. Thus we rerank the top 100 docs for each dataset. We show in Appendix C and Table 10 that our observations hold up to 10,000 documents.

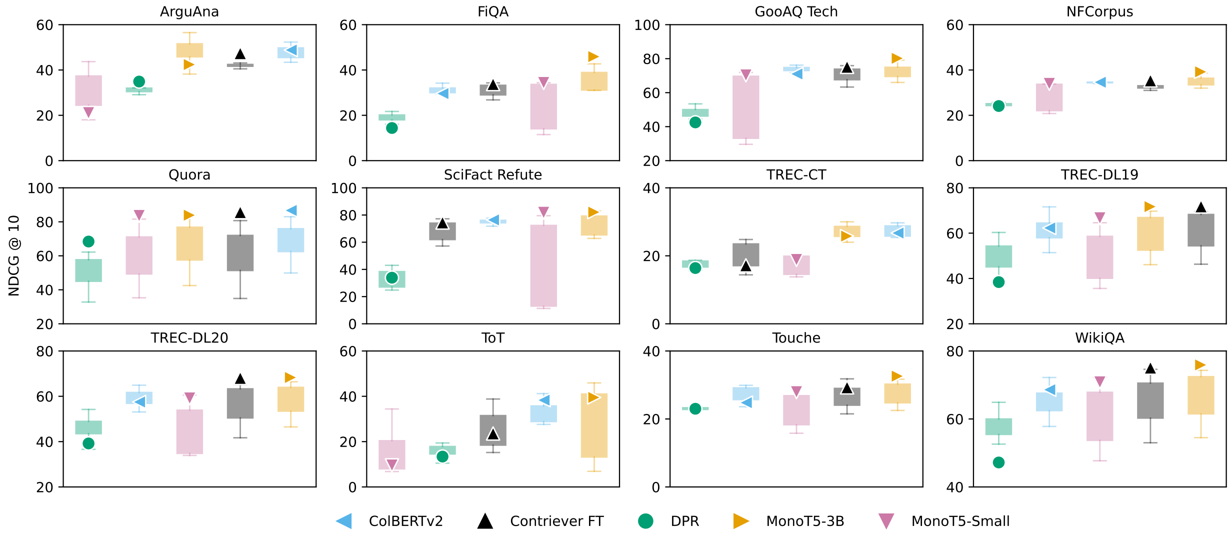

Figure 2: Effect of expansion over twelve datasets. For each dataset, markers show base performance for models, while the boxplot indicates the range of changes in scores for document and/or query expansion. Across all datasets and models, we note a consistent trend: models with **lower base performance benefit** from expansion; **higher performing rankers** generally **suffer** when expansion techniques are used.

Type	Model	DL Track 2019				FiQA				Arguana			
		No Exp	QE	DE	Both	*No Exp*	QE	DE	Both	*No Exp*	QE	DE	Both
First Stage	DPR	38.4	+6.6	+3.1	+10.8	14.4	+4.7	+1.7	+5.7	34.9	-7.1	+1.6	-4.4
	Contriever	49.0	+3.5	+4.0	+8.1	21.3	+3.6	+1.6	+5.1	45.8	-0.1	+2.9	-3.2
	BM25	51.2	-4.0	-	-	23.6	+4.5	-	-	30.0	-5.4	-	-
	Contriever FT	62.3	+1.6	-0.2	+0.6	29.6	+3.2	+0.6	+3.8	48.8	-3.6	+2.0	-2.5
	E5 Base v2	67.3	-3.4	-0.9	-3.7	37.8	-0.6	-3.8	-2.5	51.1	-8.4	+2.6	-5.7
	MPNet Base v2	68.3	-6.0	-2.9	-6.8	44.5	-4.1	-3.5	-5.7	47.6	-5.1	+5.3	-0.7
	E5 Small v2	69.1	-4.8	-1.9	-6.8	36.4	+0.4	-2.9	-0.6	46.1	-8.7	+2.7	-9.8
	GTE Large	70.0	-4.5	-1.3	-4.5	41.2	-2.0	-4.1	-3.2	56.8	-8.8	-0.9	-9.0
	E5 Large v2	70.1	-5.7	-1.7	-7.6	38.6	-0.9	-2.7	-3.2	48.9	-5.9	+3.2	-3.4
Rerankers	MonoT5-Small	66.6	-2.0	-2.8	-2.8	34.3	+0.1	-0.6	-0.3	21.1	+22.7	-3.0	+22.2
	MiniLM-2-v2	68.0	-3.2	-4.1	-5.1	27.5	-2.0	+0.6	-15.8	15.2	+11.4	+10.8	+11.2
	SPLADEv2	70.1	-4.3	-3.7	-5.6	33.4	+1.3	-0.2	+1.2	45.0	-4.5	-1.3	-4.0
	MonoBERT	70.4	-4.6	-2.0	-4.8	36.2	+0.2	-0.7	+0.0	50.1	-5.7	+2.5	-9.3
	MiniLM-4-v2	70.6	-3.0	-2.5	-4.9	33.8	+1.5	-0.3	+1.2	43.4	+0.4	+1.0	-0.8
	MonoT5-Base	71.5	-3.2	-1.4	-5.2	39.2	-1.2	-1.2	-0.9	27.0	+20.0	+0.7	+18.7
	MonoT5-3B	71.7	-2.8	-2.0	-5.0	45.9	-3.8	-3.2	-5.6	42.4	+6.8	-1.9	+5.2
	ColBERTv2	71.8	-4.2	-2.8	-6.4	33.8	-0.4	-0.3	-0.7	47.4	-5.2	-0.6	-4.8
	MiniLM-12-v2	72.0	-4.3	-4.5	-5.6	35.5	-0.4	-0.5	+0.0	33.2	+12.0	+1.1	+9.8
	MonoT5-Large	72.2	-4.0	-1.8	-5.6	42.8	-2.3	-2.3	-3.1	31.2	+14.8	-2.0	+14.8
	LLAMA	72.6	-2.9	-4.9	-7.7	40.0	-3.7	-4.9	-5.8	52.6	-3.9	-6.9	-9.4
	LLAMAv2	72.8	-4.2	-4.9	-9.3	41.1	-3.6	-7.4	-7.9	52.3	-1.5	-8.2	-7.0
	LLAMAv2-13B	73.6	-4.5	-5.4	-7.3	41.2	-4.5	-4.9	-7.0	49.4	-2.1	-6.0	-4.9

Table 1: Best expansion strategies across different models. *QE* stands for query expansion (Q-LM PRF), *DE* for document expansion (Doc2Query), and *Both* for the combination (Q-LM PRF + Doc2Query). Colors indicate a positive or negative delta over scores for no expansion. Models with higher base scores are generally harmed by expansions while weaker models benefit from them. Llama models follow MonoT5 in fine-tuning on MSMarco.

original query. Many works have shown the effectiveness of this approach (Jagerman et al., 2023; He et al., 2022; Trivedi et al., 2022).

LM-based Pseudo Relevance Feedback (Q-LM PRF). PRF is a classical IR method to expand a query using terms from top retrieved documents. We use an LM to generate a list of terms from the top 3 documents ranked by a bi-encoder model (Contriever). Through a second invocation, the LM updates the query to include the new terms. LM-aided PRF has been shown to be broadly effective (Mackie et al., 2023; Jagerman et al., 2023).

2.2 Document Expansion

Doc2Query. There are fewer widespread LM document expansion techniques, with the main one being Doc2Query (Nogueira et al., 2019c). Work has found that improving the question generation model results in higher scores, hence we use ChatGPT instead of T5 for our experiments (Nogueira et al., 2019a). See Appendix A for results using alternative LMs for document expansion.

LM-based Document PRF (D-LM PRF). Similar to the Q-LM PRF technique above, we propose

Axis	Dataset	# Queries	# Docs	Avg. Judged/Q	Q Len	D Len
In-Domain	TREC DL Track 2019 (Craswell et al., 2020)	43	8,841,823	212.5	5.4	56.6
	TREC DL Track 2020 (Craswell et al., 2021)	54	8,841,823	207.9	6.0	56.6
Domain Shift	FiQA-2018 (Maia et al., 2018)	648	57,600	2.6	10.9	137.4
	Gooaq Technical (Khashabi et al., 2021)	1,000	4,086	1.0	8.3	44.5
	NFCorpus (Boteva et al., 2016)	323	3,633	38.2	3.3	233.5
Relevance Shift	Touché-2020 (Bondarenko et al., 2020)	49	382,545	19.0	6.6	293.7
	SciFact Refute (Wadden et al., 2020)	64	5,183	1.2	12.1	214.8
Long Query Shift	Tip of My Tongue (Lin et al., 2023)	2,272	1,877	1.0	144.3	100.5
	TREC Clinical Trials '21 (Roberts et al., 2021)	75	375,580	348.8	133.3	919.5
	ArguAna (Wachsmuth et al., 2018)	1,406	8,674	1.0	197.1	170.3
Short Doc Shift	WikiQA (Yang et al., 2015)	369	26,196	1.2	6.3	25.1
	Quora (Iyer et al., 2017)	10,000	522,931	1.6	9.5	12.5

Table 2: Statistics of datasets in this work. Avg. Judged/Q is the number of relevant documents per query. Length is measured in words. The TREC DL Track uses the MS MARCO dataset (Nguyen et al., 2016).

		DL 2019 Track			DL 2020 Track		
Type	Model	DPR	Contriever FT	MonoT5-3B	DPR	Contriever FT	MonoT5-3B
–	*No Expansion*	38.4	62.3	71.7	39.2	57.5	68.3
Query	HyDE	+18.8	**+9.3**	-4.0	+13.2	**+7.4**	-5.8
	CoT	+12.6	+2.7	-6.7	+5.5	+4.2	-9.3
	Q-LM PRF	+6.6	+1.6	-2.2	+6.3	+2.7	-3.0
Doc	D2Q	+3.1	-0.2	**-1.2**	+3.1	+1.3	**-1.9**
	D-LM PRF	-1.1	-15.5	-23.6	-2.6	-9.1	-19.3
Both	HyDE + D2Q	**+21.9**	+9.0	-4.5	**+15.0**	+6.2	-5.4
	CoT + D2Q	+15.1	+0.8	-7.3	+7.2	+4.2	-8.1
	Q-LM PRF + D2Q	+10.8	+0.6	-4.2	+8.1	+3.7	-3.3
	HyDE + D-LM PRF	+16.7	-3.1	-22.8	+11.4	+1.2	-17.9
	CoT + D-LM PRF	+10.9	-10.9	-25.0	+4.1	-4.4	-21.8
	Q+D LM PRF	+6.8	-5.6	-14.4	+4.5	-2.4	-11.8

Table 3: In-Domain performance on the TREC Deep Learning Tracks, according to various types of expansions, showing that expansion typically helps weaker models (like DPR) but hurts stronger models (especially large reranker models like MonoT5-3B). Colors indicate a positive or negative delta over scores for no expansion.

a document expansion that draws pseudo-relevance from *related queries* instead of related documents. In this setting, where there exists a set of unjudged user queries, we show the LM the top 5 most-similar queries and ask it to expand the original document to better answer the relevant queries.

3 RQ1: How Do Different Models Impact Query and Document Expansion?

Experimental Setting To understand the efficacy of LM-based expansions, we employ a wide variety of neural retrieval models: DPR (Karpukhin et al., 2020); ColBERT v2 (Santhanam et al., 2022); SPLADE v2 (Formal et al., 2021a); MonoBERT (Nogueira et al., 2019b); several MonoT5 (Nogueira et al., 2020), E5 (Wang et al., 2022b), and MiniLM models (Wang et al., 2020); GTE (Li et al., 2023); `all-mpnet-v2-base` (Reimers and Gurevych, 2019); Llama 1 & 2 models (Touvron et al., 2023a,b), which we fine-tune on MS MARCO.

Due to the exponential combination of models and datasets, we evaluate all models on three representative datasets in Table 1 (we provide a comprehensive description of all datasets in §5); then, we use five representative models (DPR, Contriever, ColBERTv2, MonoT5-small, and MonoT5-3B) on a larger suite of datasets (see Figure 2).

We present results for expansion technique as absolute increase/decrease in nDCG@10[5] points over a baseline with no expansion, which we highlight in grey in all tables. Values above zero (e.g. greater than the base version) are highlighted blue while values below the base are highlighted red. Color intensity is scaled linearly according to the

[5]Traditional expansion techniques increase recall of retrieval systems. However, LM-based expansions have been shown to also improve precision (Jagerman et al., 2023). Thus, we use the official, precision-oriented metric for BEIR, nDCG.

Type	Model	FiQA-2018 DPR	FiQA-2018 Contriever FT	FiQA-2018 MonoT5-3B	GooAQ Technical DPR	GooAQ Technical Contriever FT	GooAQ Technical MonoT5-3B	NFCorpus DPR	NFCorpus Contriever FT	NFCorpus MonoT5-3B
	No Expansion	14.4	29.6	45.9	42.5	71.0	80.2	24.1	34.6	39.2
Query	HyDE	+3.6	-0.3	-14.7	+3.1	+3.8	-10.0	+0.3	+0.0	-5.9
	CoT	+3.6	+0.4	-13.2	+2.0	+2.1	-9.7	-0.7	-0.6	-4.5
	Q-LM PRF	+4.7	+3.2	-3.8	+6.4	+1.9	-3.4	+0.2	-0.4	-2.7
Doc	D2Q	+1.7	+0.6	**-3.2**	+6.4	+3.0	**-1.1**	+1.3	**+0.6**	**-0.5**
	D-LM PRF	+3.3	+1.6	-12.5	+3.8	+0.6	-11.4	+0.3	-0.3	-0.7
Both	HyDE + D2Q	+4.5	+0.4	-14.8	+8.2	**+5.2**	-7.4	**+1.6**	+0.1	-7.2
	CoT + D2Q	+4.4	+0.2	-13.4	+7.2	+3.8	-6.9	+0.8	+0.0	-5.6
	Q-LM PRF + D2Q	+5.7	+3.8	-5.6	**+10.9**	+4.2	-4.1	+1.4	-0.1	-3.0
	HyDE + D-LM PRF	+5.8	+1.2	-14.8	+5.3	+2.7	-14.2	+0.8	+0.1	-6.3
	CoT + D-LM PRF	+6.2	+1.7	-14.9	+3.6	+1.9	-13.6	-0.1	-0.2	-4.2
	Q+D LM PRF	**+7.3**	**+4.6**	-8.4	+7.9	+3.5	-6.4	+0.2	+0.0	-2.8

Table 4: How different expansions affect results on datasets that measure **Domain Shift**. Colors indicate a positive or negative delta over scores for no expansion. Notice that models with higher base scores are generally harmed by expansions while weaker models benefit from them.

Type	Model	Touche-2020 DPR	Touche-2020 Contriever FT	Touche-2020 MonoT5-3B	Scifact-Refute DPR	Scifact-Refute Contriever FT	Scifact-Refute MonoT5-3B
	No Expansion	23.0	24.8	32.6	33.9	76.4	82.1
Query	HyDE	-0.3	+4.8	-5.9	-9.1	-0.9	-12.3
	CoT	+0.3	**+5.1**	-7.4	-7.6	+0.3	-8.8
	Q-LM PRF	**+0.6**	+3.9	-1.3	+6.5	+1.1	-1.7
Doc	D2Q	-0.2	+0.0	**-0.9**	+2.0	-1.8	**+0.9**
	D-LM PRF	-0.2	-1.2	-8.3	+2.5	-4.6	-16.5
Both	HyDE + D2Q	-0.1	+5.0	-3.0	-6.1	-1.0	-16.6
	CoT + D2Q	+0.3	+2.6	-5.4	-6.5	-1.1	-16.9
	Q-LM PRF + D2Q	-0.1	+1.0	-2.0	**+9.1**	**+1.3**	-1.1
	HyDE + D-LM PRF	+0.5	+1.4	-10.1	-5.2	-2.9	-17.6
	CoT + D-LM PRF	-0.2	+0.8	-8.4	-7.2	-1.5	-19.3
	Q+D LM PRF	+0.3	+2.5	-2.7	+7.6	-2.5	-4.0

Table 5: How different expansions affect results on datasets that measure **Relevance Shift**.

difference between the base value and the min/max (*i.e.*, more saturation for the highest/lowest values).

We use default hyperparameters for all models, except for the length of the queries, which we set at 512 for BERT-based models and 1024 for T5 and Llama models.

Effect of Different Models Our results with all models (Figure 1) show a consistent pattern: as base performance on a task increases, the gains from expansion decrease. We also see this trend from Table 1 (note that ArguAna and FIQA results are sorted by nDCG score on MS MARCO; negative trend is clearly observable in Figure 1). Interestingly, these results do not depend on the model architecture: this is true for bi-encoders, late-interaction models, neural sparse models, and cross-encoders (of all sizes).

However, do these results hold for other datasets? In Figure 2, we show that this pattern is consistent over a wide range of datasets. Models whose base score is higher (such as MonoT5-3B) are negatively impacted by expansions.

4 RQ2: How Do Different Distribution Shifts Impact Results?

Experimental Setting We evaluate how query and document expansion are impacted by different distribution shifts: in-domain/no shift (MS MARCO), domain shift (e.g. medical, code, legal), relevance shift (finding the opposite or a counterargument), and format shift (extremely long queries or very short documents). Datasets and their descriptive statistics are in Table 2. We use three representative models for these experiments.

In-Domain We use two datasets that test performance on the MS MARCO collection: TREC Deep Learning[6] 2019 and 2020 tracks (Craswell et al.,

[6]Despite the different names, TREC DL 2019 and 2020 use the same document collection as MS MARCO, albeit with new queries and relevance judgements.

Type	Model	Tip of My Tongue DPR	Tip of My Tongue Contriever FT	Tip of My Tongue MonoT5-3B	TREC CT 2021 DPR	TREC CT 2021 Contriever FT	TREC CT 2021 MonoT5-3B	Arguana DPR	Arguana Contriever FT	Arguana MonoT5-3B
	No Expansion	13.4	38.3	39.5	16.4	26.7	25.8	34.9	48.8	42.4
Query	HyDE	+3.0	-9.4	-26.8	+0.3	+2.1	**+4.2**	-4.5	-5.4	**+15.8**
	CoT	+2.1	-9.5	-23.3	**+2.3**	**+3.0**	+3.0	-5.8	-5.3	+11.3
	Q-LM PRF	-2.9	-1.9	**+6.4**	+2.2	+0.6	-0.1	-7.1	-3.6	+8.3
Doc	D2Q	+1.6	-3.2	-8.5	+0.3	-1.3	-1.8	+1.6	+2.0	-2.1
	D-LM PRF	+5.5	**+2.9**	+0.9	-0.7	-0.9	+0.6	**+2.3**	**+3.5**	-2.5
Both	HyDE + D2Q	+3.6	-10.7	-29.7	+0.4	+2.1	+2.7	-2.8	-2.5	+12.9
	CoT + D2Q	+2.2	-10.6	-25.3	**+2.3**	+1.5	-0.1	-4.3	-3.0	+10.6
	Q-LM PRF + D2Q	-1.8	-4.7	+2.1	+0.7	-0.9	-0.2	-4.4	-2.5	+6.9
	HyDE + D-LM PRF	**+6.0**	-7.2	-32.6	+0.0	+1.0	+3.2	-3.0	+1.0	+10.3
	CoT + D-LM PRF	+5.3	-7.4	-25.8	+1.9	+2.7	+1.0	-4.0	+0.9	+8.8
	Q+D LM PRF	+0.7	+1.6	**+6.4**	+0.6	-1.0	+0.4	-4.0	-0.2	+3.3

Table 6: How different expansions affect results on datasets that measure **Long Query Format Shift**. Colors indicate a positive or negative delta over scores for no expansion. *Unlike previous results*, all models benefit from expansions on all three datasets. We conclude that, in the case of significant query shift, expansion is useful.

Type	Model	WikiQA DPR	WikiQA Contriever FT	WikiQA MonoT5-3B	Quora DPR	Quora Contriever FT	Quora MonoT5-3B
	No Expansion	47.2	68.6	75.9	68.4	86.7	83.9
Query	HyDE	+16.4	**+3.6**	**-1.6**	-15.4	-13.8	-8.2
	CoT	+9.8	-0.9	-6.1	-32.3	-31.5	-35.4
	Q-LM PRF	+11.9	-2.2	-4.2	-13.8	-11.4	-7.0
Doc	D2Q	+5.4	-1.8	-1.7	**-6.2**	**-3.7**	**+0.0**
	D-LM PRF	-2.8	-10.8	-21.4	-10.0	-15.6	-17.0
Both	HyDE + D2Q	**+17.7**	+2.1	-2.7	-11.4	-10.1	-7.1
	CoT + D2Q	+11.3	-1.5	-6.9	-25.7	-26.3	-32.5
	Q-LM PRF + D2Q	+13.0	-1.1	-6.2	-9.4	-8.7	-6.9
	HyDE + D-LM PRF	+12.6	-6.2	-18.0	-21.1	-22.1	-20.2
	CoT + D-LM PRF	+7.0	-10.3	-19.0	-35.6	-36.8	-41.4
	Q+D LM PRF	+9.5	-6.1	-10.8	-19.4	-19.6	-17.8

Table 7: How different expansions affect results on datasets that measure **Short Document Format Shift**. Models with higher base scores are generally harmed by expansions while weaker models benefit from them.

2020, 2021). All retrieval models considered train on MS MARCO, hence these are *in-domain*.

Domain Shift In this setting models must generalize from training domain (web documents from MS MARCO) to new domains, such as legal or medical text. This type of shift is made difficult by specialized vocabulary in these domains. We use NFCorpus (medical) (Boteva et al., 2016), GooAQ Technical (code) (Khashabi et al., 2021), and FiQA-2018 (finance) (Maia et al., 2018).

Relevance Shift This setting is characterized by a difference in how *relevance* is defined. Rather than topical relevance over web pages, queries in these datasets ask for counterarguments or documents refuting its claim. We use two datasets that search for refutations or counterarguments: Touché-2020 (Bondarenko et al., 2020) and a subset of SciFact (Wadden et al., 2020) whose gold documents refute the queries claims.

Format Shift Another type of shift is the length of inputs: generally, queries are short and documents span over one to multiple paragraphs. However, there are situations where queries could be document-sized or the documents could be short. This shift tests whether models can generalize to new length formats. We consider two sets of datasets: for *shift to long query* we use the "Tip of My Tongue" dataset introduced by Lin et al. (2023), TREC Clinical Trials Track 2021 (Roberts et al., 2021), and ArguAna (Wachsmuth et al., 2018). For *shift to short document*, we use Quora (Iyer et al., 2017) and WikiQA (Yang et al., 2015).

4.1 Results by Type of Shift

Table 3 shows results for in-domain data on the 2019 and 2020 Deep Learning TREC Tracks. We see that weaker models improve with different expansion types, with DPR improving for almost every expansion and the stronger Contriever showing

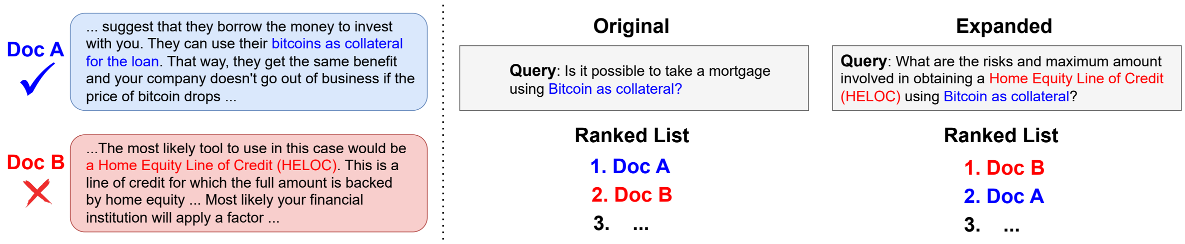

Figure 3: An example of expansions obscuring the relevance signal. The non-relevant document in red (×) was ranked higher than the relevant blue (✓) document due to the phrase *"Home Equity Line of Credit"* being added to the query. The left side shows the original query and documents while the right side shows the ranking.

minor improvements for some combinations. However, when we move to the stronger models (*e.g.*, MonoT5-3B), we find that all of these gains disappear and expansions hurt the model.

We find that this trend holds in most other categories of shift: Table 4 for domain shift, Table 5 for relevance shift, and Table 7 for short document shift. Note that Figure 2 also shows this visually.

The exceptions to this pattern occur in format shift: on Quora (Table 5), all models are harmed by expansion; for long query shift (Table 6), expansions generally help most models. When we examine why expansions help for the latter, we find that the transformations typically shorten queries to more closely resemble models' training data (*e.g.*, for ArguAna the query changes from a long document to a shorter sentence that summarizes it).

As IR models are not typically trained on long queries, it is an open-question of whether additional training would make this category of shift easier for models and thus make expansions less helpful.

5 RQ3: Why Do Expansions Hurt?

Sections 3 and 4 show that strong IR models do not benefit from expansions. But what causes this effect? Here, we explore whether model size (§5.1) is linked to our findings, and perform a qualitative error analysis (§5.2).

5.1 Drop in Performance Independent of Size

One possible argument is that larger models are able to estimate relevance better when using unaltered queries and documents, as they have learned a more refined relevance model during their training. To verify this hypothesis, we test two different families of models: MonoT5 and E5. If model size is the cause, we would expect to see larger models gain less from expansions for both families.

However, Figure 5 shows that model scale is inversely correlated with gains from expansion for the MonoT5-family, but not the E5-family. The crucial difference between them[7] can be attributed to the E5 models having similar performance scores across sizes whereas T5 has a much wider range: T5 differs by 21 nDCG@10 points on ArguAna from 3B to small while E5 differs by only 3 points from large to small. Thus, we see that model size impacts gains from expansions only in tandem with the correlation between model size and base score.

5.2 Error Analysis

If model size is not the reason for our finding, what could be causing it? To gain an intuition on the failures of LM expansion, we annotate 30 examples from three datasets where performance declines when expanding queries and documents.

We find that out of the 30 examples, two are false negatives, *i.e.*, relevant documents that are unjudged and not labeled as relevant (both from FiQA). Of the remaining 28, all errors are due to the expansions adding irrelevant terms that dilute relevance signal, or including erroneous keywords that make irrelevant documents appear relevant. Figure 3 shows an example of how query expansion added the term *"Home Equity Line of Credit"* and distracted from the main focus of the question (using bitcoins as collateral). Thus, it is likely that, without the noise LM-based expansions introduce, well tuned rankers can accurately estimate relevance of subtly different documents. We can visualize this in Figure 4, where we note a general downward shift of the rankings of relevant documents in the top-10 positions for TREC DL 2019. We find that most expansions shifts the ranking by a few positions, while some expansions shift the relevant document ranks to be out of the top 10 (i.e. the cluster at -10 in Figure 4).

[7]Another obvious difference is that E5 is a bi-encoder while MonoT5 is not. However, previous work (Muennighoff, 2022) has shown that bi-encoders also improve with scale.

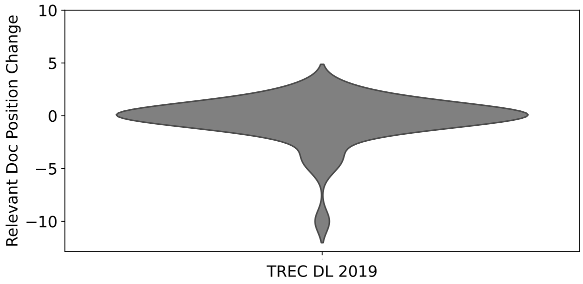

Figure 4: The change in rank for relevant documents in the top 10 when using expansions. Negative values indicate lower ranks (e.g. -5 indicates that the rank of the relevant document went down 5 when using expansions). We see that expansions cause relevant documents to be ranked lower. Figure 6 in the Appendix shows other datasets with similar results.

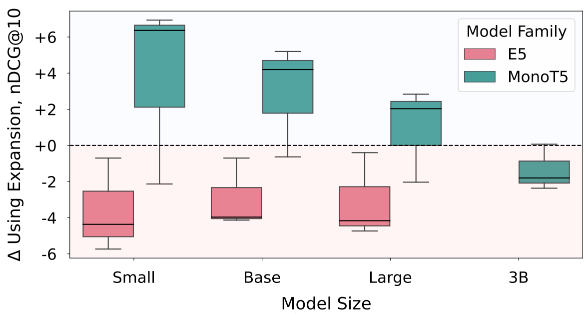

Figure 5: Model scale does **not** explain negative effect of LM-based expansions. While larger MonoT5 models perform worse, all E5 model sizes are equally impacted

6 Discussion

Our results indicate three phenomena regarding expansion using LMs: (*i*) expansion generally benefits weaker models, such as DPR, while better performing rankers, such as T5, are penalized; (*ii*) exceptions are observed in case of severe distribution shift, *e.g.* very long queries; (*iii*) when model scores decrease, the cause is generally expansion weakening the original relevance signal.

This implies that despite their broad capabilities, LMs should not be used to augment strong performing IR models without careful testing. The strong performance of rerankers for generalization confirms previous work by Rosa et al. (2022a). Further, Table 3 indicates this characterization of LM expansion also holds on in-domain data (no shift).

Interestingly, our experiments find that the only distribution shift that consistently needs expansion is long query format shift; we found no equivalent result for domain, document, or relevance shift. Future work may examine whether improved training techniques on longer queries can overcome this or whether longer queries are innately more difficult.

7 Related Work

Large Scale Analyses in Neural IR Comprehensive analyses in retrieval have provided great insight into practical uses of retrieval. These include many aspects of information retrieval, including interpretability (MacAvaney et al., 2022), domain changes (Lupart et al., 2023), syntax phenomena (Chari et al., 2023; Weller et al., 2023), and relationship between neural and classical IR approaches (Formal et al., 2021b; Chen et al., 2022).

Generalization in Neural IR As retrieval models have become more effective, attention has turned to improving and evaluating the way that IR models generalize to out-of-distribution datasets (e.g. not MS MARCO-like corpora). One prominent example of this is the BEIR dataset suite (Thakur et al., 2021), which is commonly used for retrieval evaluation. Much other work has proposed new datasets for types of shift (e.g. MTEB (Muennighoff et al., 2023) among others (Han et al., 2023; Ravfogel et al., 2023; Weller et al., 2023)), as well as many new modeling strategies for better zero-shot retrieval (Dai et al., 2022; Wang et al., 2022a). We follow these works by showing different types of shift and whether these types of shift change the results for LM-based expansion techniques.

Effect of Scale on Neural IR Models IR models typically improve with scale (Nogueira et al., 2020) but are also heavily constrained, due to the requirement of processing documents for live search. Thus, most first-stage IR models typically use a BERT backbone (Santhanam et al., 2022; Izacard et al., 2021) while reranker models have scaled to billions of parameters (Nogueira et al., 2020). However, work on scaling bi-encoder architectures has also shown performance gains from scale (Muennighoff, 2022). Due to the effectiveness of larger models, recent work has shown that a better first-stage model does not lead to improvements over a BM25 + reranker pipeline (Rosa et al., 2022a). Thus, for our experiments we use BM25 as first stage retrieval and show results reranking those.

Query and Document Expansion in IR Query and document expansion have a long history in IR, with early techniques such as expanding query terms using dictionaries or other hand-built knowledge sources (Smeaton et al., 1995; Liu et al., 2004) as well as techniques that use corpus-specific information such as pseudo-relevance feedback (Rocchio Jr, 1971). These expansions are limited as they are either hand-crafted (and thus limited in scope) or involved automatic techniques that may introduce spurious connections between words. LM-based query and document expansions on the other hand can rely on their extensive linguistic knowledge which goes well beyond hand-crafted rules. Despite this however, they still suffer from spurious and superfluous additions, as shown in Figure 3. However, LM-based expansions have been shown to be successful in a variety of applications (Zheng et al., 2020; Weller et al., 2022; Wang et al., 2023a; Jagerman et al., 2023), which provided inspiration for this work.

8 Conclusion

We conduct the first large scale analysis on large language model (LM) based query and document expansion, studying how model performance, architecture, and size affects these results. We find that these expansions improve weaker IR models while generally harming performance for the strongest models (including large rerankers and heavily optimized first-stage models). We further show that this negative correlation between model performance and gains from expansion are true for a wide variety of out of distribution datasets, except for long query shift, where this correlation is weaker. Overall, our results indicate that LM expansion should not be used for stronger IR models and should instead be confined to weaker retrieval models.

Limitations

We evaluate rankers in a zero-shot setup. This work does not train rankers to deal with augmentations. While additional training might help mitigate negative effect of document and query expansion, it would significantly increase computational requirements. In fact, as our analysis reveals that no single expansion technique is superior in all settings, users would need to train rankers for multiple expansion techniques, further increasing the cost of this fine-tuning step. Finally, some tasks might require fine-tuning on supervised data, which might not be available or easily obtainable.

Our protocol for choosing whether a ranker need expansion requires labeled test data in the target domains. While our work requires no labeled data to train models, we note that deciding whether to use augmentation requires having access to evaluation data for the target domain: in some cases, such data might not be available. While recently proposed LM-aided IR evaluation techniques (Faggioli et al., 2023; MacAvaney and Soldaini, 2023; Thomas et al., 2023) might ameliorate the need of supervised data, we do not explore such approaches in this work.

While open LMs were evaluated, majority of experiments rely on commercial LM APIs. The majority of experiments in this work were carried out with commercial language models available via paid APIs. While we experimented with a variety of other paid API and open LMs (`gpt-4-0613`, Claude V2, Llama2 70b Chat), we found that they all generally show similar trends, with commercial APIs currently outperforming open models (see Appendix A and Table 8 for more details). As our work is mainly focused on studying the effect of expansion different rankers, we feel picking one representative model is justified. Nevertheless, use of commercial APIs limits reproducibility and presents a significant barrier to those who cannot get access to the model. To minimize this, we will release all LM generations gathered from commercial APIs and from open-source models.

Compute requirements to fully replicate this work. A replication of this work would require access to significant computational resources, including GPUs. A rough estimate shows that generating results for this paper required north of 10,000 NVIDIA A6000 GPU hours, with a further 5,000 hours to develop a stable experimental platform.

Only English information retrieval models are evaluated. This work only studies datasets in English. While LM augmentations could play an important role in improving non-English, cross-lingual, and multilingual information retrieval, they require careful analysis (c.f. Mayfield et al. (2023) as one example).

Ethical Considerations

LMs may generate factually incorrect text, which could affect ranking. This work shows that LM augmentations make mistakes; while our experimental setup is such that LM-generated content never replaces actual documents, inaccuracies might result in non-relevant documents being presented to users.

Acknowledgements

OW is supported by the National Science Foundation Graduate Research Fellowship Program. We thanks Sean MacAvaney, Akari Asai, and the Semantic Scholar team at AI2 for their feedback and comments that helped improve this work, as well as those of the reviewers.

References

Nasreen Abdul-Jaleel, James Allan, W Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Mark D Smucker, and Courtney Wade. 2004. Umass at trec 2004: Novelty and hard. *Computer Science Department Faculty Publication Series*, page 189.

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. 2022. Task-aware retrieval with instructions. *ArXiv preprint*, abs/2211.09260.

Jagdev Bhogal, Andrew MacFarlane, and Peter Smith. 2007. A review of ontology based query expansion. *Information processing & management*, 43(4):866–886.

Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2020. Overview of Touché 2020: Argument Retrieval. In *Working Notes Papers of the CLEF 2020 Evaluation Labs*, volume 2696 of *CEUR Workshop Proceedings*.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *Proceedings of the 38th European Conference on Information Retrieval (ECIR 2016)*, pages 716–722.

Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1).

Andreas Chari, Sean MacAvaney, and Iadh Ounis. 2023. On the effects of regional spelling conventions in retrieval models. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Xilun Chen, Kushal Lakhotia, Barlas Oguz, Anchit Gupta, Patrick Lewis, Stan Peshterliev, Yashar Mehdad, Sonal Gupta, and Wen-tau Yih. 2022. Salient phrase aware dense retrieval: Can a dense retriever imitate a sparse one? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 250–262, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the trec 2020 deep learning track.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the trec 2019 deep learning track. *ArXiv preprint*, abs/2003.07820.

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 preprint*, abs/2209.11755.

Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2023. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '23, page 39–50, New York, NY, USA. Association for Computing Machinery.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021a. Splade v2: Sparse lexical and expansion model for information retrieval. *ArXiv preprint*, abs/2109.10086.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021b. Match your words! a study of lexical matching in neural information retrieval. In *European Conference on Information Retrieval*.

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

Rujun Han, Peng Qi, Yuhao Zhang, Lan Liu, Juliette Burger, William Yang Wang, Zhiheng Huang, Bing Xiang, and Dan Roth. 2023. Robustqa: Benchmarking the robustness of domain adaptation for open-domain question answering. In *Annual Meeting of the Association for Computational Linguistics*.

Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*.

Armin Hust, Stefan Klink, Markus Junker, and Andreas Dengel. 2003. Towards collaborative information retrieval: Three approaches.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. Quora question pairs. *First Quora Dataset Release: Question Pairs*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *ArXiv preprint*, abs/2112.09118.

Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. *ArXiv preprint*, abs/2305.03653.

Vitor Jeronymo, Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, Jakub Zavrel, and Rodrigo Nogueira. 2023. Inpars-v2: Large language models as efficient dataset generators for information retrieval. *ArXiv preprint*, abs/2301.01820.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen Tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 6769–6781. Association for Computational Linguistics (ACL).

Daniel Khashabi, Amos Ng, Tushar Khot, Ashish Sabharwal, Hannaneh Hajishirzi, and Chris Callison-Burch. 2021. GooAQ: Open question answering with diverse answer types. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 421–433, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Victor Lavrenko and W. Bruce Croft. 2001. Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, page 120–127, New York, NY, USA. Association for Computing Machinery.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *ArXiv preprint*, abs/2308.03281.

Kevin Lin, Kyle Lo, Joseph E Gonzalez, and Dan Klein. 2023. Decomposing complex queries for tip-of-the-tongue retrieval. *ArXiv preprint*, abs/2305.15053.

Shuang Liu, Fang Liu, Clement T. Yu, and Weiyi Meng. 2004. An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Simon Lupart, Thibault Formal, and Stéphane Clinchant. 2023. Ms-shift: An analysis of ms marco distribution shifts on neural retrieval. In *Advances in Information Retrieval*, pages 636–652, Cham. Springer Nature Switzerland.

Sean MacAvaney, Sergey Feldman, Nazli Goharian, Doug Downey, and Arman Cohan. 2022. ABNIRML: Analyzing the behavior of neural IR models. *Transactions of the Association for Computational Linguistics*, 10:224–239.

Sean MacAvaney and Luca Soldaini. 2023. One-shot labeling for automatic relevance estimation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2230–2235, New York, NY, USA. Association for Computing Machinery.

Iain Mackie, Shubham Chatterjee, and Jeffrey Stephen Dalton. 2023. Generative relevance feedback with large language models. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

James Mayfield, Eugene Yang, Dawn Lawrie, Samuel Barham, Orion Weller, Marc Mason, Suraj Nair, and Scott Miller. 2023. Synthetic cross-language information retrieval training data. *arXiv preprint arXiv:2305.00331*.

Niklas Muennighoff. 2022. Sgpt: Gpt sentence embeddings for semantic search. *ArXiv preprint*, abs/2202.08904.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.

Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019a. From doc2query to docttttquery. *Online preprint*, 6:2.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019b. Multi-stage document ranking with bert. *ArXiv preprint*, abs/1910.14424.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019c. Document expansion by query prediction. *ArXiv preprint*, abs/1904.08375.

Shauli Ravfogel, Valentina Pyatkin, Amir D. N. Cohen, Avshalom Manevich, and Yoav Goldberg. 2023. Retrieving texts based on abstract descriptions. *ArXiv preprint*, abs/2305.12517.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, Steven Bedrick, and Willian R Hersh. 2021. Overview of the trec 2021 clinical trials track. In *Proceedings of the Thirtieth Text REtrieval Conference (TREC 2021)*.

Joseph John Rocchio Jr. 1971. Relevance feedback in information retrieval. *The SMART retrieval system: experiments in automatic document processing*.

Guilherme Rosa, Luiz Bonifacio, Vitor Jeronymo, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022a. In defense of cross-encoders for zero-shot retrieval. *ArXiv preprint*, abs/2212.06121.

Guilherme Moraes Rosa, Luiz Henrique Bonifacio, Vitor Jeronymo, Hugo Abonizio, Marzieh Fadaee, Roberto de Alencar Lotufo, and Rodrigo Nogueira. 2022b. No parameter left behind: How distillation and model size affect zero-shot retrieval. *ArXiv preprint*, abs/2206.02873.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Alan F Smeaton, Fergus Kelledy, and Ruairi O'Donnell. 1995. Trec-4 experiments at dublin city university: Thresholding posting lists, query expansion with wordnet and pos tagging of spanish. *Harman [6]*, pages 373–389.

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

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2023. Large language models can accurately predict searcher preferences. *ArXiv*, abs/2309.10621.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *ArXiv preprint*, abs/2212.10509.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022a. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2345–2360, Seattle, United States. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022b. Text embeddings by weakly-supervised contrastive pre-training. *ArXiv preprint*, abs/2212.03533.

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

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Xiao Wang, Sean MacAvaney, Craig Macdonald, and Iadh Ounis. 2023b. Effective contrastive weighting for dense query expansion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12688–12704.

Xiao Wang, Craig Macdonald, Nicola Tonellotto, and Iadh Ounis. 2023c. Colbert-prf: Semantic pseudo-relevance feedback for dense passage and document retrieval. *ACM Transactions on the Web*, 17(1):1–39.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Orion Weller, Aleem Khan, Nathaniel Weir, Dawn Lawrie, and Benjamin Van Durme. 2022. Defending against misinformation attacks in open-domain question answering. *arXiv preprint arXiv:2212.10002*.

Orion Weller, Dawn J Lawrie, and Benjamin Van Durme. 2023. Nevir: Negation in neural information retrieval. *ArXiv preprint*, abs/2305.07614.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: BERT and beyond. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, pages 1–4, Online. Association for Computational Linguistics.

Zhi Zheng, Kai Hui, Ben He, Xianpei Han, Le Sun, and Andrew Yates. 2020. Bert-qe: contextualized query expansion for document re-ranking. *arXiv preprint arXiv:2009.07258*.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji rong Wen. 2023. Large language models for information retrieval: A survey. *ArXiv preprint*, abs/2308.07107.

A Different LMs for Expansion

Here we show results for GPT-4 expansions instead of ChatGPT in Table 8. We can see that although absolute numbers differ slightly, there is no change to the trends discussed in the main paper: i.e. that stronger models are harmed by expansions while weaker models benefit. We swap FiQA for NFCorpus due to the larger collection size and increased costs of annotating with GPT-4.

B Placement of Expansions

In Table 9 we show different placements of expansions (i.e. do we prepend/append/replace the original query when doing query expansion?). We find that the placement does not make a significant difference to our overall results, as the core conclusion of the paper remains the same.

C Number of Documents Reranked

All experiments reranked the top 100 documents, as described in the main paper. However, here we show an ablation where we increase the number of reranked documents to include the top 1000 or 10,000 documents. We see that there is little change in how expansions affect the models.

D Example Input/Output for expansions

Here we show example input and output for expansion for query and document expansions.

D.1 Document Expansion

We show example expansion from this document from MS MARCO:

> Whale Shark. The Whale Shark is the largest and the heaviest living shark because its weight is more than 21 tons and its length is more than 12 meter. It is found in open oceans and warm waters. It feeds on plankton but some times it also uses other small fishes as food.

D.1.1 Doc2Query

> What is the largest and heaviest living shark in the world?
> What are the characteristics of the Whale Shark?
> Where can the Whale Shark be found?
> What does the Whale Shark feed on?
> What is the size and weight of the Whale Shark?

D.1.2 D-LM PRF

> The Whale Shark is the largest and heaviest living shark, weighing over 21 tons and measuring over 12 meters in length. It is found in open oceans and warm waters. While it primarily feeds on plankton, it occasionally consumes other small fishes as well.

D.2 Query Expansion

For the MS MARCO query “why did the us volunterilay enter ww1" the following output was generated. For prompt formatting, see Appendix E.

Type	Model	TREC DL 2019 DPR	TREC DL 2019 Contriever FT	TREC DL 2019 MonoT5-3B	NFCorpus DPR	NFCorpus Contriever FT	NFCorpus MonoT5-3B	Arguana DPR	Arguana Contriever FT	Arguana MonoT5-3B
	No Expansion	38.4	62.3	71.7	24.1	34.6	39.2	34.9	48.8	42.4
ChatGPT	Q-LM PRF	+6.6	**+1.6**	-2.8	+0.2	-0.4	-2.8	-7.1	-3.6	**+6.8**
ChatGPT	D2Q	+3.1	-0.2	**-2.0**	+1.3	**+0.6**	**-0.5**	**+1.6**	**+2.0**	-1.9
ChatGPT	Q-LM PRF + D2Q	**+10.8**	+0.6	-5.0	**+1.4**	-0.1	-3.0	-4.4	-2.5	+5.2
GPT-4	Q-LM PRF	**+13.3**	**+5.2**	**-0.6**	-7.8	-17.5	-22.6	-6.2	-4.5	+4.5
GPT-4	D2Q	-4.3	-14.0	-2.3	**+1.2**	**+1.0**	**-0.1**	**+0.9**	**+1.2**	+0.2
GPT-4	Q-LM PRF + D2Q	+8.0	-8.6	-3.2	-7.6	-17.8	-23.3	-4.8	-2.9	**+5.2**
Claude v2	PRF	+14.0	**+4.8**	-3.7	+0.3	+1.1	-1.5	-6.0	-5.7	**+4.0**
Claude v2	D2Q	+4.2	-1.7	**-2.4**	**+1.6**	+0.5	**-0.2**	**+3.4**	**+3.3**	-1.0
Claude v2	PRF + D2Q	**+15.3**	+2.6	-4.4	+1.5	**+1.6**	-1.6	-3.1	-2.1	+3.7
Llama v2 70B Chat	PRF	+0.9	-8.3	-14.5	-1.5	-1.7	-3.9	-4.8	-4.5	-2.6
Llama v2 70B Chat	D2Q	**+4.7**	**-1.1**	**-2.5**	**+1.0**	**+0.2**	**-0.2**	**-0.1**	**+0.9**	**-2.5**
Llama v2 70B Chat	PRF + D2Q	+3.6	-7.8	-15.8	-0.7	-1.7	-4.2	-4.3	-3.4	-4.0

Table 8: How different LLMs used as the generator affect results. Colors indicate a positive or negative delta over scores for no expansion. Although there are small differences **the overall trends are the same**.

Type	Model	MSMarco 2019 Contriever	MSMarco 2019 MonoT5-small	MSMarco 2019 MonoT5-3B	FiQA Contriever	FiQA MonoT5-small	FiQA MonoT5-3B	Arguana Contriever	Arguana MonoT5-small	Arguana MonoT5-3B
	No Expansion	14.4	29.6	45.9	42.5	71.0	80.2	24.1	34.6	39.2
Query	Prepend	+8.1	-2.8	-4.2	+5.1	-0.3	-5.6	-3.2	+22.2	+6.9
Query	Append	+9.8	-1.6	-3.5	+4.1	**+0.8**	-4.6	-3.5	+22.6	**+8.4**
Query	Replace	+8.3	-7.3	-7.9	+7.2	-3.2	-8.8	-15.9	+19.3	+3.3
Doc	Prepend	+8.5	-2.2	-1.9	+5.9	-2.0	-3.1	+1.4	-5.4	-12.4
Doc	Append	+10.3	-0.8	-1.4	+4.0	-1.4	-2.2	+0.4	-6.8	-8.6
Doc	Replace	+9.3	-8.9	-6.2	**+8.3**	-6.9	-8.8	-4.1	-11.0	-20.1
Both	Prepend/Prepend	+9.4	-2.2	-2.0	+5.9	-4.0	-4.6	+1.5	-9.7	-19.8
Both	Prepend/Append	**+11.0**	-0.9	-1.9	+4.1	-3.3	-2.8	+0.5	-8.7	-18.3
Both	Prepend/Replace	+9.6	-9.0	-6.2	+8.1	-8.5	-9.3	-5.1	-10.0	-26.8
Both	Append/Prenpend	+3.5	-2.0	-2.2	+3.6	+0.1	-3.8	-0.1	**+22.7**	+8.3
Both	Append/Append	+2.7	-1.7	-1.1	+4.8	-3.5	-2.0	-0.5	-5.3	-9.0
Both	Append/Replace	+3.0	-1.7	-1.3	+4.6	-5.6	-2.2	-0.3	-8.0	-18.8
Both	Replace/Prepend	+4.0	-2.8	-1.2	+1.6	-0.6	-3.2	**+2.9**	-3.0	-2.1
Both	Replace/Append	+5.9	**+0.2**	**-0.7**	+0.9	+0.6	**-1.2**	+1.2	-1.5	-0.9
Both	Replace/Replace	+5.7	-11.8	-8.7	+4.4	-5.3	-10.4	-1.0	-5.0	-9.1

Table 9: How different placements of the expansions affect results (e.g. prepend/append/replace). Colors indicate a positive or negative delta over scores for no expansion. Although there are small differences **the overall trends are the same**.

Type	Model	TREC DL 2019 DPR	TREC DL 2019 Contriever FT	TREC DL 2019 MonoT5-3B	NFCorpus DPR	NFCorpus Contriever FT	NFCorpus MonoT5-3B	Arguana DPR	Arguana Contriever FT	Arguana MonoT5-3B
100 Docs	*No Expansion*	38.4	62.3	71.7	24.1	34.6	39.2	34.9	48.8	42.4
100 Docs	Q-LM PRF	+6.6	**+1.6**	-2.8	+0.2	-0.4	-2.8	-7.1	-3.6	**+6.8**
100 Docs	D2Q	+3.1	-0.2	**-2.0**	+1.3	**+0.6**	**-0.5**	**+1.6**	**+2.0**	-1.9
100 Docs	Q-LM PRF + D2Q	**+10.8**	+0.6	-5.0	**+1.4**	-0.1	-3.0	-4.4	-2.5	+5.2
1k docs	*No Expansion*	29.2	64.6	72.6	21.5	34.2	40.0	29.5	48.7	38.0
1k docs	PRF	+4.1	-0.6	-4.2	-0.9	+0.1	-3.5	-10.1	-14.5	**-1.9**
1k docs	D2Q	+5.2	**+0.7**	**-1.7**	**+1.1**	**+0.7**	**-0.3**	**-5.2**	**-11.3**	-10.4
1k docs	PRF + D2Q	**+10.9**	-0.1	-6.3	+0.7	+0.6	-3.9	-8.8	-13.5	-2.8
10k docs	*No Expansion*	28.5	64.6	73.1	21.5	34.2	40.0	23.5	36.2	29.7
10k docs	PRF	+3.4	+0.6	-4.3	-0.7	**+0.6**	-3.2	-3.9	-1.9	**+6.6**
10k docs	D2Q	+4.2	-1.4	**-1.5**	**+0.9**	+0.5	**-0.2**	**+1.2**	**+1.6**	-2.2
10k docs	PRF + D2Q	**+10.1**	**+2.8**	-4.2	+0.3	**+0.6**	-3.4	-2.1	-1.0	+5.8

Table 10: How different numbers of documents used for reranking affect the results. Colors indicate a positive or negative delta over scores for no expansion. Although there are small differences between models **the overall trends are the same**. Note that NFCorpus's collection is roughly 3k examples and so changing from 1k to 10k reranking makes little difference.

D.2.1 HyDE

> The United States voluntarily entered World War I primarily due to a combination of factors. Firstly, the sinking of the British ocean liner RMS Lusitania by a German U-boat in 1915, resulting in the loss of American lives, created a strong public outcry for intervention. Additionally, Germany's resumption of unrestricted submarine warfare in 1917 further escalated tensions. Furthermore, the interception of the Zimmermann Telegram, which revealed Germany's proposal to Mexico for a military alliance against the United States, heightened fears of potential threats to American security. These events, coupled with a desire to protect American interests and uphold democratic principles, ultimately led the United States to voluntarily join the war effort.

D.2.2 Q-LM PRF

> What were the factors leading to the United States' voluntary entry into WW1 and what were the consequences of their decision to join the Allies?

D.2.3 Chain of Thought

> The United States voluntarily entered World War I primarily due to a combination of factors. One of the key reasons was the German unrestricted submarine warfare, which threatened American lives and interests at sea. Additionally, the interception and decoding of the Zimmerman Telegram, in which Germany proposed an alliance with Mexico against the US, further fueled public sentiment for intervention. Moreover, there was a desire to protect American economic interests, as the US had significant trade ties with the Allied powers. Lastly, President Woodrow Wilson's belief in promoting democracy and his vision for a post-war world order based on the principles of self-determination and collective security also played a role in the US decision to enter the war.

E Prompts for LM expansions

Here we show the prompts used for LM query and document expansion.

E.1 Query Expansions

E.1.1 HyDE

We use the prompts from their paper, which are dataset specific, in this manner:

> <SPECIFIC PROMPT>
>
> <Type of Input>: QUERY
> <TYPE OF OUTPUT> (one short paragraph max):

for example on Clinical Trials we use:

> Please write a clinical trial summary that would apply to the following patient.
>
> Patient Info: QUERY
> Trial Summary (one short paragraph max):

and on FiQA we use:

> Please write a financial article passage to answer the question
>
> Question: QUERY
> Passage (one short paragraph max):

E.1.2 Q-LM PRF

> You are a query expansion engine, primed and ready to take in text and output additional keywords will provide new and expanded context behind the original input. Your extensive world knowledge and linguistic creativity enables you to provide questions that maximally optimize the new questions to find new websites. You **always** provide creative synonyms and acronym expansions in your new queries that will provide additional insight.
>
> Be sure to use new words and spell out acronyms (or add new acronyms). Hint: think of ***new synonyms and/or acronyms*** for “QUESTION" using these documents for inspiration:
>
> DOCUMENTS
>
> Return the following information, filling it in:
> Input: QUESTION
> Comma Separated List of 10 important New Keywords: “““NEW KEYWORDS HERE"""
> New Question (combining Input and New Keywords, only **one** new question that expands upon the Input): “““NEW QUESTION HERE"""
>
> Your output:

E.1.3 Chain of Thought

We use a the same specific prompt for CoT as we do for HyDE. The format is as follows:

> <SPECIFIC PROMPT>
>
> QUESTION
>
> Give the rationale (one short paragraph max) before answering.

E.2 Document Expansions

E.2.1 D-LM PRF

> Change the following document to answer these questions, if they are partially answered by the document. If the queries are not relevant, ignore them. Your new documents should be one concise paragraph following the examples.
>
> Example 1:
>
> Queries:
> 1. “how much caffeine is in a 12 ounce cup of coffee?"
> 2. “what are the effects of alcohol and caffeine"
> 3. “what can pregnant women not do?"
>
> Document: “We don’t know a lot about the effects of caffeine during pregnancy on you and your baby. So it’s best to limit the amount you get each day. If you are pregnant, limit caffeine to 200 milligrams each day. This is about the amount in 1½ 8-ounce cups of coffee or one 12-ounce cup of coffee."
> New Document (similar to Document): “There is a lack of research about the effects of caffeine during pregnancy on you and your baby. So it’s best to limit the amount you get each day. If you are pregnant, limit caffeine to 200 milligrams (mg) each day. This is about the amount in 1½ 8-ounce cups of coffee or one 12-ounce cup of coffee (e.g. 200 milligrams)."
>
> Example 2:
>
> Queries:
> QUERIES
> Document: “DOCUMENT"
> New Document (similar to Document):

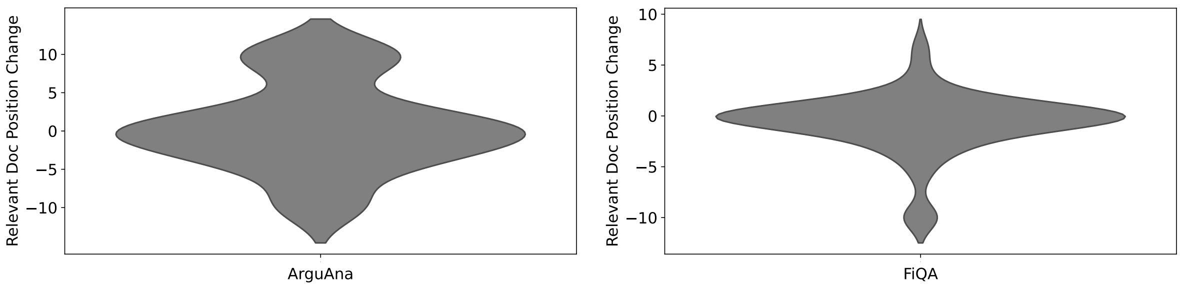

Figure 6: Number of positions relevant documents change when using expansion. Negative values indicate the document was ranked lower. Results are similar to TREC DL 2019 for FiQA which shows lowered nDCG while for Arguana nDCG scores increase as seen by the change in positions being positive.

E.2.2 Doc2Query

You are an optimized query expansion model, ExpansionGPT. You will write 5 queries for the given document that help retrieval models better find this document during search.

Document: “QUESTION"

Queries: