Few-shot Prompting for Pairwise Ranking: An Effective Non-Parametric Retrieval Model

Nilanjan Sinhababu
Centre for Computational and Data Sciences
IIT Kharagpur, India
nilanjansb@kgpian.iitkgp.ac.in

Andrew Parry
School of Computing Science
University of Glasgow
United Kingdom
a.parry.1@research.gla.ac.uk

Debasis Ganguly
School of Computing Science
University of Glasgow
United Kingdom
Debasis.Ganguly@glasgow.ac.uk

Debasis Samanta
Department of Computer Science and Engineering
IIT Kharagpur, India
dsamanta@iitkgp.ac.in

Pabitra Mitra
Department of Computer Science and Engineering
IIT Kharagpur, India
pabitra@cse.iitkgp.ac.in

Abstract

A supervised ranking model, despite its effectiveness over traditional approaches, usually involves complex processing - typically multiple stages of task-specific pre-training and fine-tuning. This has motivated researchers to explore simpler pipelines leveraging large language models (LLMs) that can work in a zero-shot manner. However, since zero-shot inference does not make use of a training set of pairs of queries and their relevant documents, its performance is mostly worse than that of supervised models, which are trained on such example pairs. Motivated by the existing findings that training examples generally improve zero-shot performance, in our work, we explore if this also applies to ranking models. More specifically, given a query and a pair of documents, the preference prediction task is improved by augmenting examples of preferences for similar queries from a training set. Our proposed pairwise few-shot ranker demonstrates consistent improvements over the zero-shot baseline on both in-domain (TREC DL) and out-domain (BEIR subset) retrieval benchmarks. Our method also achieves a close performance to that of a supervised model without requiring any complex training pipeline.

1 Introduction

Development of novel neural architectures and training methodologies (Pradeep et al., 2021; Izacard et al., 2022a; Formal et al., 2021; Wang et al., 2023; Karpukhin et al., 2020a) have substantially outperformed the unsupervised approaches. Commonly, these neural approaches involve deep interactions between embedded representations of queries and documents (Dai and Callan, 2019; Khattab and Zaharia, 2020) thus overcoming the vocabulary mismatch problem of discrete term representations. However, to achieve good performance, not only do these deep neural models require a large number of training data in the form of pairs of example queries and their relevant documents (Gao and Callan, 2022; Saeed et al., 2021), but the effectiveness of these models also depends on a number of ad-hoc decision choices, e.g., the following.

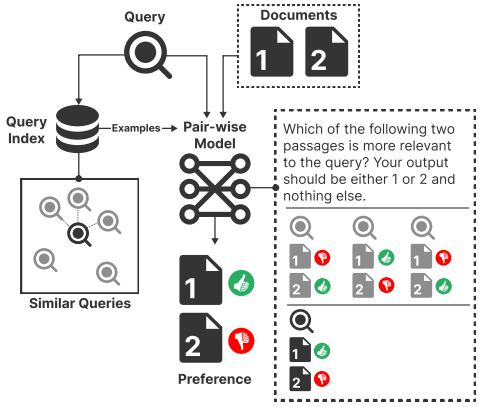


Figure 1: Our proposed pairwise method for reranking a set of top-retrieved candidate documents via LLM-based inference. Different from Qin et al. (2023), we provide additional context for LLM inference by including few-shot examples, each consisting of documents relevant to queries similar to the current input query as retrieved from a training set.

Findings of the Association for Computational Linguistics: EACL 2024, pages 12363–12377

- *Neural Architecture*, e.g., bi-encoder (Karpukhin et al., 2020b), cross-encoder (Nogueira and Cho, 2019) or learned sparse models (MacAvaney et al., 2020);
- *Pre-Training Tasks* - to effectively capture retrieval specific term semantics (Gao and Callan, 2021; Gao et al., 2021);
- *Index construction* and the *number of ranking stages*, e.g., sparse retrieval followed by reranking (Pradeep et al., 2021; Lassance et al., 2024), vs. approximate inner product search on a dense vector index (Lin et al., 2020; Izacard et al., 2022a),
- *Negative Selection* - methodology for noise contrastive learning objective (Xiong et al., 2020; Hofstätter et al., 2021; Cohen et al., 2024);
- *Training Data Augmentation* via generative models (Bonifacio et al., 2022; Dai et al., 2023);
- *Distillation Strategies* - for effectively transferring the representational capabilities of larger models into smaller ones (Xiao et al., 2022; Lin et al., 2020);
- *Curriculum Selection* - i.e., the order in which training samples are presented, for knowledge distillation (He et al., 2023; Zeng et al., 2022).

With such a large number of decision choices available, it is difficult to converge on a set of 'best practices' for designing the pipeline of a supervised ranker. Instead, in this paper, we present a relatively simple approach of leveraging the information from a training set of examples of query and relevant document pairs *without requiring any parametric training*. The motivation for this simple yet effective pipeline stems from the recent developments in large language models (LLMs), which exhibit emergent capabilities of modeling semantics (Ma et al., 2023b). By including a task definition and, optionally, examples of labelled or unlabelled data, these models can perform competitively on unseen tasks without fine-tuning (Sun et al., 2023; Qin et al., 2023). As such, they offer a compelling alternative to the highly data-driven fine-tuning currently applied in the field of neural retrieval.

Although recent work has employed LLMs for ranking, these approaches either simply use a zero-shot approach, thus not leveraging the benefits of a non-parametric memory (i.e., of a training set) (Qin et al., 2023), or they employ zero-shot inference only as an initial step for training data augmentation prior to train a supervised ranking model (Zhuang et al., 2022; Ouyang et al., 2022a). In contrast, our work, while on the one hand, employs an unsupervised approach (i.e., with no parametric training involved), it is able to make use of the training data of query-relevance examples via few-shot prompting on the other.

More specifically, as outlined in Figure 1, in our approach, following the methodology of (Qin et al., 2023) we input a query and a pair of documents seeking to estimate the relative preferential order between the pair. However, in contrast to the PRP (pairwise rank prompting) method of Qin et al. (2023), we input a set of queries (from an available training set) that are related to the current query in terms of an abstract similarity measure. This is motivated from the well-known cognitive bias *attribute substitution* effect in psychology, where to answer an unknown question, human brains *recollect known answers to related questions* and eventually *process information from them to answer the unknown one* (Kahneman and Frederick, 2002; Honda et al., 2017).

In our work, we emulate this behavior of attribute substitution heuristics on LLMs, where in additional context (Brown et al., 2020), we provide examples of related queries and their relevant documents. The hypothesis is that LLMs, with their inherent language processing abilities, should be able to make a more informed judgment of pairwise relevance by processing the examples provided.

This proposed workflow requires investigating a number of research questions and challenges including: "how to retrieve an effective set of related queries for a given input?", and "what is the downstream effect of the similarity of the information needs of the related queries". Our experiments show that an embedding-based neighborhood to retrieve related queries yields better downstream effectiveness than a lexical model, and we also show that a small number of examples can, in fact, lead to consistent improvements over the zero-shot PRP approach (Qin et al., 2023). Our experiments also demonstrate the potential of this unsupervised method for out-of-domain generalization. More specifically, we show that examples queries from MS-MARCO lead to improvements on TREC Covid and SciFact test collections. The source code of our proposed LLM-based few-shot ranker[1] is made available for research purposes.

[1] https://github.com/nilanjansb/fewshot_prp

2 Related Work

Zero-shot Information Retrieval with LLMs. Generative language models exhibit generalization capabilities beyond the tasks on which they are trained (Ouyang et al., 2022b; Touvron et al., 2023). Naturally, this has led to a number of effective zero-shot approaches to NLP tasks. Sun et al. (2023) first proposed the use of LLMs as cross-encoders in a list-wise ranking setting. Similar results were observed by Ma et al. (2023c) using a different prompting strategy. A common thread of work proposes distilling list-wise closed source models into smaller decoder-only architectures (Pradeep et al., 2023a,b; Zhang et al., 2023). They report that, in some cases, a student model could outperform a significantly larger teacher model (Pradeep et al., 2023b). Beyond list-wise ranking, LLMs have additionally been applied in a bi-encoder setting (Ma et al., 2023a).

Qin et al. (2023) first applied large language models to pair-wise ranking, finding that in a truly zero-shot setting, such an approach was competitive on out-of-domain benchmarks. Zhuang et al. (2024a) further improved both the efficiency and effectiveness of pair-wise ranking.

In-Context Learning (ICL). In-context learning or few-shot learning is an inference strategy that differs from the standard notion of supervised learning in the sense that labeled examples are appended to a model instruction improving effectiveness in an out-of-domain downstream task (Ni et al., 2021; Li et al., 2022). Though initially considered to guide sequence generation in tasks such as question answering and abstractive summarization (Li et al., 2023; Tang et al., 2023), ICL has been shown to be effective in classification-style tasks (Lu et al., 2022; Milios et al., 2023) and, therefore, could be effective in a cross-encoder setting for ranking. In terms of example selection for ICL, prior work has found that conditioning chosen examples on the current test instance is effective (Nie et al., 2022; Xie et al., 2023).

3 Methodology

3.1 Overview of Zero-shot PRP

Pointwise Relevance Score Estimation. A parameterized pointwise ranking model, given a query Q and a candidate document D involves computing a relevance estimation function of the form $\mathcal{S}(Q, D; \theta)$, where θ denotes a parameter vector trained by noise contrastive loss in a pairwise (Pradeep et al., 2021) or listwise manner (Zhuang et al., 2022; Sun et al., 2023). As candidates for the pairwise comparisons, it is common to employ a standard sparse retrieval (e.g., BM25) and then compute the likelihood values for each pair.

Unlike a supervised approach, which involves optimizing the parameters θ of a model via pairwise or listwise loss functions, an LLM-based ranker employs its frozen parameters to predict the relevance score. In the simplest possible setting, this takes the form of pointwise predictions, i.e., $\mathcal{S}(Q, D; \theta) = f(Q, D, \theta_{\text{LLM}})$, where θ_{LLM} refers to frozen pre-trained parameters (not fine-tuned specifically with a ranking objective). In practice, the function $f(Q, D, \theta_{\text{LLM}})$ represents a function of the posterior probability (logits) of a pre-specified set of tokens, e.g., the function f for Mono-T5 is defined as $e^{\theta(\text{`true'})}/(e^{\theta(\text{`true'})} + e^{\theta(\text{`false'})})$.

Pairwise Rank Prompting (PRP). Although such pointwise relevance score estimation has been used for training data augmentation (Sun et al., 2023), IR system evaluation (Faggioli et al., 2024) and also for query performance prediction (Meng et al., 2024), Qin et al. (2023) has shown that for the purpose of ranking, pairwise estimation of relevance is more effective than the simpler pointwise approach. More specifically, instead of explicitly predicting $\mathcal{S}_{\theta_{\text{LLM}}} : Q, D \mapsto \mathbb{R}$, an LLM decoder is now used to predict the relative preference order between a pair of documents D and D'. Formally, the prediction is of the form

$$f(Q, D, D', \theta_{\text{LLM}}) \mapsto \mathbb{I}(D \succ D'), \qquad (1)$$

where $D \succ D'$ indicates that it is more likely that D is more relevant to the query Q than D', meaning that D should be *preferred* over D'.

In practice, to estimate the preference score of a pivot document D against another document D', two different predictions are obtained from an LLM with two different input prompts - first with the sequence (D, D') and the second with the order swapped, i.e., (D', D). More specifically, the probability that the first document in the sequence is to be preferred over the second one is given by $\theta_{1,2} = e^{\theta(\text{`1'})}/(e^{\theta(\text{`1'})} + e^{\theta(\text{`2'})})$, and the complementary probability $\theta_{2,1}$ is given by swapping '1' with '2' in the expression. If these two probabilities are consistent, i.e., both $\theta_{1,2}(D, D') > \theta_{2,1}(D, D')$ and $\theta_{2,1}(D', D) > \theta_{1,2}(D', D)$ are true then the

preference score of D against D' is set to 1. Similarly, the preference score of D against D' is set to 0 for the other consistent alternative. The score is set to an uncertainty level of $1/2$ for inconsistent predictions. In a compact notation, the preference score of a pivot document D with respect to another document D' is thus defined as

$$P(D \succ D') = \frac{1}{2}[\mathbb{I}(\theta_{1,2}(D, D') > \theta_{2,1}(D, D')) + \mathbb{I}(\theta_{2,1}(D', D) > \theta_{1,2}(D', D))]. \quad (2)$$

Clearly, $P(D \succ D') \in \{0, \frac{1}{2}, 1\}$.

Finally, to obtain the overall score of a single document D, a common practice in pairwise inference models (Pradeep et al., 2021; Qin et al., 2023) is to aggregate the relative preference indicators of a pivot document D against every other document D' in the top-k retrieved candidate set $\mathcal{D}_k$, i.e.,

$$\mathcal{S}(Q, D, \theta_{\text{LLM}}) = \sum_{D' \in \mathcal{D}_k - \{D\}} P(D \succ D'), \quad (3)$$

with $P(D \succ D')$ as defined in Equation 2.

3.2 Proposed Few-shot PRP

Utilising Training Queries. The estimated preference indicators of Equation 1 depend only on the text of the current query (Q), and that of the document pairs (D and D'). Therefore, unlike a supervised model, Equation 1 is unable to make use of information from a training set of query-relevance example pairs of the form $\mathcal{Q} = \cup_i (Q_i, \mathcal{R}(Q_i))$.

We propose to modify Equation 1 by making the LLM generation process depend also on an additional context of the relevance/non-relevance information from training set queries that are similar to the input query Q. More formally, the k-shot version of the function f is now defined as

$$f_k(Q, D, D', \mathcal{N}_k(Q), \theta_{\text{LLM}}) \mapsto \mathbb{I}(D \succ D'), \quad (4)$$

where $\mathcal{N}_k(Q)$ indicates a neighborhood of k similar queries from a training set $\mathcal{Q}$, i.e.,

$$\mathcal{N}_k(Q) = \cup_{i=1}^{k} \{Q' \in \mathcal{Q} : Q' = \arg\max_i \sigma(Q, Q')\}, \quad (5)$$

where the notation $\arg\max_i$ indicates the index of the i^{th} largest value, and $\sigma(Q, Q')$ denotes a generic similarity measure between the query pair (Q, Q'). As practical choices for the query similarity function $\sigma(Q, Q')$, we employ a lexical (BM25) and a semantics-based approach (BERT). Although a fine-tuned supervised model, e.g., one that is trained on query-relevance semantics, can potentially yield better neighbourhoods of queries for ICL, we avoid using such models for neighbourhood construction in order to keep our approach completely unsupervised and non-parametric.

In practice, to select k-shot examples, we first construct a neighbourhood of top-K ($K > k$) candidate queries by employing a sparse or a dense index. Since the downstream effect of an example on an LLM's inference is not a deterministic function, we do not solely rely on the similarity function σ itself, i.e., BM25 or BERT. Instead, we randomly sample a subset of k examples from this set of top-K candidates.

Positives and Hard Negatives. A training set query $Q' \in \mathcal{Q}$ contains examples of relevant documents $\mathcal{R}(Q')$. For each query $Q' \in \mathcal{Q}$, we sample a single relevant document $R_{Q'} \sim \mathcal{R}(Q')$. In addition, following the common practice of noise contrastive learning (Xiong et al., 2020), we sample a non-relevant document as a hard negative from ranks m to M (m < M) of a BM25 retrieved list of documents for the training query Q', i.e., $N_{Q'} \sim \mathcal{D}_M(Q') - \mathcal{D}_m(Q') : N_{Q'} \notin \mathcal{R}(Q')$. Specifically, for our experiments $m = 100$ and $M = 200$, and $\mathcal{D}_k(Q)$ denotes the top-k BM25 list of documents for a query Q.

The triple $\langle Q', R_{Q'}, N_{Q'} \rangle$ constitutes a single example that we input to an LLM. To avoid the bias of setting the ground-truth preference indicator label to always a '1', we randomly flip the pair to $(N_{Q'}, R_{Q'})$, in which case the reference label becomes '2' (see Figure 2). We then repeat the process until k examples are included.

The post-inference process is identical to that of Equations 2 and 3, the only difference being that the relative preference scores now depend on the additional context of the examples and the reference preference indicators of these examples.

4 Experiment Setup

Research Questions. Since our proposed methodology is a relatively simple way to leverage information from a training set of query-relevance example pairs, the first research question is directed towards finding if this additional context from a training set helps improve zero-shot performance. Explicitly stated,

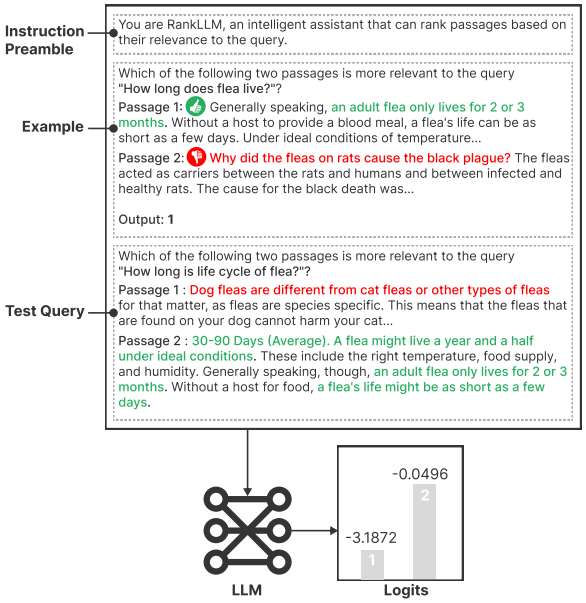


Figure 2: An example prompt to illustrate the structure of the prompts used for few-shot PRP.

- **RQ-1**: Does leveraging information from example query-relevance pairs improve retrieval effectiveness over zero-shot PRP?

Nie et al. (2022) has shown that a localized neighbourhood of examples similar to the current instance helps improve the performance of ICL. In our proposed approach (Equation 4), as particular choices for the query neighbourhood selection function $\mathcal{N}_k(Q)$, we use BM25 (sparse BoW) and BERT (dense). Both the neighbourhood functions aim to retrieve queries from a training set that potentially has an information need that is similar to the current input query. More precisely,

- **RQ-2**: Does employing queries with information needs that are potentially similar to the current query as few-shot examples in PRP help improve retrieval effectiveness?

Next, we explore the out-of-domain generalization of our non-parametric approach, i.e., the objective is to see if examples of relevant documents for source domain queries (that are likely to be topically shifted information needs as caused by the change of domain) can still help improve retrieval on the target domain.

- **RQ-3**: Can queries retrieved from a training set of a source domain, when used as examples in few-shot PRP, improve retrieval effectiveness on target domain queries?

Table 1: Statistics of the datasets used in our experiments. The $|\bar{Q}|$ and $|\bar{D}|$ denote the average number of query and document terms, respectively.

Coll	#Docs	Topics	Topics	$\lvert\bar{Q}\rvert$	$\lvert\bar{D}\rvert$
MS MARCO Passage	8.8M	Train	≈503K	5.97	56.11
		DL'19	43	5.40	
		DL'20	54	6.04	
BEIR	171332	TREC-COVID	50	3.48	182.25
	5183	SciFact	300	13.05	209.86

Datasets. We evaluate our approach on the MS MARCO passage collection (Bajaj et al., 2016) comprising over 8.8 million documents collated from the Bing search engine and then segmented into relatively short passages. For IR evaluation, we use the TREC deep learning track topics of 2019 and 2020, respectively denoted as DL'19 (Craswell et al., 2020) and DL'20 (Craswell et al., 2021).

To investigate RQ-3, we employ two test collections from the BEIR dataset (Thakur et al., 2021) for out-of-domain evaluation - namely the TREC Covid (Wang et al., 2020) and the SciFact collections (Wadden et al., 2020). While the former is a corpus of academic papers about COVID-19 and related coronavirus research, the latter is a scientific claim validation dataset. Table 1 summarizes the datasets used in our experiments.

4.1 LLM settings and Evaluation

Evaluation metrics. Following the standard convention (Craswell et al., 2021, 2020), for the DL topic sets, as evaluation metrics we report the mean average precision (MAP) at cut-off 100 (MAP@100) with the binary relevance threshold set to 2, and normalized discounted cumulative gain (nDCG) computed at cut-off 10 (nDCG@10). For BEIR, again following the standard practice, we report nDCG@10 (Qin et al., 2023; Thakur et al., 2021).

As a qualitative measure of the relatedness between the example queries and the input query, we compute the average topical similarity of the neighbourhood of the input query. In particular, we report a term-overlap-based similarity between the input query and its neighbours. Specifically, as the term overlap measure, we employ Jaccard similarity between query pairs (Zendel et al., 2019). Formally,

$$J(\mathcal{N}_k(Q)) = \frac{1}{k} \sum_{Q' \in \mathcal{N}_k(Q)} \frac{Q \cap Q'}{Q \cup Q'}. \quad (6)$$

For a given benchmark set of topics, we report

the average value of $J(\mathcal{N}_k(Q))$ aggregated over all queries, which we report in our results with the notation $\bar{J}(\mathcal{N}_k)$.

LLM Details. We employ the following as foundation LLMs for the 0-shot and few-shot PRP.

- **FLAN-T5**: This is an encoder-decoder model, specifically an instruction fine-tuned version of the T5 model (Chung et al., 2022). Our experiment uses the FLAN-T5-XL (3B parameters) variant. Although the model yields effective zero-shot performance as reported by Qin et al. (2023), this model is limited to an input size of 512 tokens only, which somewhat limits the number of examples for few-shot PRP.
- **Zephyr**, a decoder-only LLM, is a fine-tuned version of Mistral (7B) (Jiang et al., 2023) fine-tuned on publicly available synthetic datasets. The maximum input length of this LLM is 4096, which affords a greater number of examples for few-shot PRP.

4.2 Methods Investigated

For all the methods investigated on TREC DL topics, we employ a two-stage reranking pipeline, with BM25 as the first-stage ranker. The different second-stage rankers, which we describe next, are used to rerank the top 100 documents. Since the OOD retrieval task (on BEIR) is primarily a precision-centric one, we rerank only the top 20 documents from BM25, the first-stage ranker.

Baselines. We compare our few-shot PRP with the following baselines.

- **BM25**: A strong term-weighting approach operating over bag-of-words representations.
- **0-shot PRP** (Qin et al., 2023): This is the zero-shot PRP methodology outlined in Equation 1. As foundation models, we employ both Flan-T5, which was also used by Qin et al. (2023), and Zephyr. As a naming convention, we append the suffix '**0S**' to the underlying LLM's name, e.g., '**Zephyr-0S**'. As a score computation strategy, we used aggregation over all pairs of documents in the top-k set, as prescribed by Qin et al. (2023) and Pradeep et al. (2021).
- **Few-shot PRP with static examples**: This baseline uses static few-shot examples (i.e., the same example across all test queries) instead of a local neighborhood of related queries for a given input query. The objective of this baseline is to confirm if topically related contexts indeed help improve the ranking task, as is usually the case for other NLP tasks such as text classification (Liu et al., 2022). Similar to the zero-shot PRP baseline, we refer to this method with the name of the LLM used followed by the suffix k**S**, e.g., '**Zephyr-1S**'.

In addition to the above baselines, we report results with an effective supervised cross-encoder **monoT5**. It is a T5 model fine-tuned on the MS MARCO training queries in a point-wise setting using the probability of the token 'true' as an estimate of relevance. Although an unsupervised PRP approach is not directly comparable to a supervised approach, such as monoT5, we nonetheless include the results of an effective supervised model as a reference point for comparison. While presenting the results in Table 4, to prevent direct comparisons between unsupervised PRP and supervised monoT5 models, we gray out the latter.

Variants of proposed method. We employ two methodologies for the neighbourhood selection function (Equation 4) to obtain localized few-shot examples. In particular, we index the collection of MS MARCO training set queries and then employ BM25 and a dense index of the [CLS] pooled BERT vectors to obtain the candidate top-k. In the existing naming convention, for BM25, we add the suffix '**LEX**', whereas '**SEM**' is the suffix for the embedded vector-based approach. For instance, '**Zephyr-LEX-1S**' indicates 1-shot PRP, with BM25 being the similarity function to retrieve the top matching candidate. As argued in Section 3.2, to add non-determinism to the process of example selection, we sample the top-k ($k < 10$) candidates from a neighbourhood of size $K = 10$.

As an ablation, we employ a 'relevant-document' only (i.e., without the negatives) few-shot approach to observe if using only the relevant document is sufficient. This requires modifying the prompt such that the example triple we input to an LLM becomes a pair $\langle Q', R_{Q'} \rangle$. We add the suffix '**RO**' to indicate a relevant-document-only few-shot. For instance, '**Zephyr-LEX-1S-RO**' is a 1-shot PRP with BM25 similarity function and uses only the relevant documents as ICL examples.

5 Results

5.1 Main observations

Examples significantly improve retrieval effectiveness. In answering **RQ-1**, it can be observed

Table 2: A comparison between the 0-shot PRP (Qin et al., 2023) and our few-shot extension to it with two different neighborhood similarity functions to retrieve the examples. Each one-shot result reported in this table is an average over 5 runs with the standard deviations included in superscript. The best scores across all unsupervised approaches are bold-faced, and the overall best results are both bold-faced and underlined. Letters a to d are used to indicate the statistical significance of a retriever with Zephyr-0S, Zephyr-LEX-1S, Zephyr-SEM-1S, and monoT5.

		TREC DL'19			TREC DL'20		
Type	Retriever	$\bar{J}(\mathcal{N}_k)$	AP@100	nDCG@10	$\bar{J}(\mathcal{N}_k)$	AP@100	nDCG@10
Baseline	BM25	n/a	.2322	.4795	n/a	.2719	.4950
	Contriever	n/a	.2910	.6346	n/a	.3776	.6292
	FLAN-T5-0S	n/a	.3431	.6574	n/a	.3654	.6184
	Zephyr-0S	n/a	.3220	.6420	n/a	.3305	.5782
Ours	FLAN-T5-LEX-1S	.267	$.3338^{(.0003)}$	$.6515^{(.0034)}$	.244	$.3720^{(.0008)}$	$.6291^{(.0050)}$
	FLAN-T5-SEM-1S	.352	$.3357^{(.0008)}$	$.6543^{(.0042)}$	.370	$.3746^{(.0020)}$	$.6284^{(.0009)}$
	Zephyr-LEX-1S	.267	$\underline{.3447}^{(.0019)a}$	$\underline{.6742}^{(.0005)a}$	.244	$\underline{.3793}^{(.0052)abc}$	$\underline{.6457}^{(.0077)abc}$
	Zephyr-SEM-1S	.352	$\mathbf{.3512}^{(.0041)a}$	$\mathbf{.6785}^{(.0028)a}$	.370	$\mathbf{.3824}^{(.0019)abc}$	$\mathbf{.6480}^{(.0033)abc}$
Ablation	FLAN-T5-1S	.041	$.3279^{(.0023)}$	$.6418^{(.0033)}$	.029	$.3733^{(.0024)}$	$.6204^{(.0019)}$
	Zephyr-1S	.041	$.3440^{(.0029)a}$	$.6697^{(.0072)a}$	.029	$.3565^{(.0026)}$	$.6001^{(.0043)}$
	Zephyr-LEX-1S-RO	.267	$.3269^{(.0009)}$	$.6390^{(.0035)}$	.244	$.3711^{(.0026)}$	$.6251^{(.0011)}$
	Zephyr-SEM-1S-RO	.352	$.3096^{(.0013)}$	$.6137^{(.0027)}$	.370	$.3444^{(.0019)}$	$.6021^{(.0021)}$
Supervised	monoT5	n/a	$\mathbf{.3570}^{a}$	$\mathbf{.6998}^{a}$	n/a	$\mathbf{.3970}^{a}$	$\mathbf{.6729}^{a}$

from Table 2 that on providing annotated pair-wise examples, retrieval effectiveness is improved in terms of nDCG@10 on both DL'19 and DL'20 test queries. Specifically, in a zero-shot setting, FLAN-T5 outperforms Zephyr. In a few-shot setting, FLAN-T5 effectiveness either degrades or improves by a small margin (0.01 on nDCG@10) showing no significant change in effectiveness. A likely reason for this ineffectiveness of Flan-T5 in a few-shot setting, as compared to Zephyr, can likely be attributed to the characteristic differences in their instruction tuning phases.

Our approach is also competitive with monoT5 (a supervised model), and is statistically indistinguishable from supervised approaches in-domain. Though we do not outperform a supervised approach, the fact that an unsupervised approach's performance is close to that of a supervised one indicates that our proposed few-shot PRP method successfully leverages the benefits of a training set of query-relevance pairs without involving the complex stages and decision choices (related to, e.g., neural architecture, negative selection, distillation strategies etc.) as typically required for a supervised ranker.

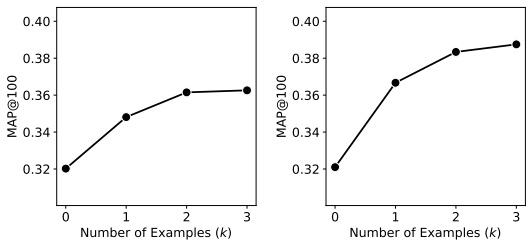
Similar queries yield effective examples. Concerning **RQ-2**, which is our core contribution in this work, we find that in considering the locality of a given annotated query to a test instance, we can further improve the effectiveness of ICL in ranking as shown in Rows 6 and 7 of Table 2. Our method also improves on a static baseline (i.e., where examples are not selected as per a similarity function but are rather chosen in a static manner).

We further explore the effects of using both lexical and semantic similarity scoring functions, for example, selection. Additionally, while few-shot PRP significantly improves over a zero-shot baseline in all cases, our ablation using static examples does not.

Due to both inverted index structures and approximate nearest neighbor indices, our approach has minimal overhead relative to random selection. Furthermore, as we select by query locality, our approach has no additional overhead incurred due to the increase in ranking depths.

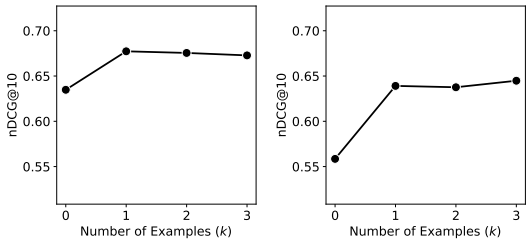
We find that in-domain selection by semantic similarity is more effective than lexical similarity, with retrieval effectiveness following a linear trend to Jaccard similarity. Much like standard retrieval a lexical model will suffer from term mismatch whereas a semantic model can find similar queries by sequence-level context.

A higher number of examples yields greater precision at lower depths. In Figure 3, we observe that MAP@100 is monotonically increasing with increasing values of k - the number of examples in few-shot PRP. The metric nDCG@10 plateaus beyond $k = 1$. We posit that given the precision-orientated nature of re-ranking, a smaller value of k may be preferable as this also saves computation time.



(a) MAP@100 on DL'19

(b) MAP@100 on DL'20



(c) nDCG@10 on DL'19

(d) nDCG@10 on DL'20

Figure 3: Sensitivity of Zephyr-kS on #few-shot examples.

Table 3: Evaluating (nDCG@10) re-ranking performance on top-20 BM25 retrieved documents in out-of-domain settings. The query-document relevance pairs are retrieved from MS MARCO to construct the ICL example sets for other test collections. Here, only the BM25 and Zephyr-0S baselines, supervised monoT5 ranker, and our localized 1S methods are compared. Letters a to d are used to indicate the statistical significance of a retriever with Zephyr-0S, Zephyr-LEX-1S, Zephyr-SEM-1S, and monoT5 (paired t-test with $p = 0.05$).

	TREC Covid		SciFact	
Retriever	$\bar{J}(\mathcal{N}_k)$	nDCG@10	$\bar{J}(\mathcal{N}_k)$	nDCG@10
BM25	n/a	.5781	n/a	.6722
Contriever	n/a	.4499	n/a	.6477
Zephyr-0S	n/a	.6571	n/a	.6872
Zephyr-LEX-1S	.130	**.6790**^{ad}	.093	**.6988**
Zephyr-SEM-1S	.094	.6753^{d}	.067	.6880
monoT5	n/a	.6376	n/a	**.7204**^{ac}

However, if using our approach in a distillation setting for annotation to deeper ranks, it may be worthwhile to increase k because, in an offline process, this overhead would be less important. A likely reason for the plateau of nDCG@10 may be due to the "lost-in-the-middle" effect (Liu et al., 2023), which points to the characteristic behaviour that decoder-only models place greater importance on the start and end of a sequence. In the context of our task, it turns out that even a single annotated example is sufficient to differentiate relevant documents from non-relevant ones, thus avoiding any "lost-in-the-middle" type effects.

In-domain examples improve effectiveness out-of-domain. Regarding **RQ-3**, in Table 3 we present results using MSMARCO annotated queries with our selection method over out-of-domain (OOD) corpora. As Zephyr was found to be more effective in a few-shot setting, we do not assess FLAN-T5 in this setting.

An important observation is that the topical overlap of the similar queries (N_k) with the current input query is much lower for OOD, e.g., '.093' for SciFact vs. '.370' as obtained with the semantic neighbourhood for in-domain evaluation on TREC DL'20 (see Table 2). This is expected as the SciFact or Covid queries cover different topics of information needs as compared to the MSMARCO training set queries.

Despite this reduced topical overlap of the few-shot examples, we observe that they are useful in improving the zero-shot performance (the few-shot approach also outperforms monoT5 on the Covid dataset). Similar to Table 2, we observe a positive correlation between the topical overlap of the related information needs and the ranking effectiveness (higher topical overlap leads to better retrieval results). This finding is important as, beyond in-domain tests, our approach shows generalisation on par with or exceeding a strong supervised model.

In summary, we have shown that not only does our method require no parametric training, but being a non-parametric approach also enables it to adapt to changing corpora. It can perform competitively against both strong unsupervised and fine-tuned retrieval models. For a task such as retrieval augmented generation (Lewis et al., 2020), our model could be used as both the 'retriever' and the 'reader'.

5.2 Qualitative analysis

Table 4 shows an example when 1-shot PRP (Zephyr-1S) can improve the rank of a relevant document from 16 to 1 thus contributing to a substantial increase in nDCG@10 value. In this case, the example query '*Which airport in Paris is closest to the city*' is largely similar to the current input query '*Is CDG airport in main Paris*', which means that the relevant document provided for the example query indeed provides useful signals to the generative process. It is likely that the underlined text segments of the example relevant document, e.g., '*potential relocations*' and '*expand the airport*' provide useful semantic cues - that the CDG airport

Table 4: An example query from DL'19, where Zephyr-SEM-1S improves the rank of a relevant document from 16 to 1.

Current query: Is CDG airport in main Paris? **Relevant document**: Paris Charles de Gaulle Airport IATA: CDG, ICAO: LFPG also known as Roissy Airport (name of the local district), is the largest international airport in France. It is named after Charles de Gaulle (1890-1970), leader of the Free French Forces during the Second World War, founder of the French Fifth Republic and President of France from 1959 to 1969. Charles de Gaulle Airport is located within portions of several communes 25 km (16 mi) to the northeast of Paris.
1-shot training query: Which airport in Paris is closest to the city? **1-shot training relevant document**: Paris Charles de Gaulle airport covers 32.38 square kilometres (12.50 sq mi) of land. The choice of this vast area was made based on the limited number of potential relocations and expropriations and the possibility to further expand the airport in the future.

is close to the main city of Paris - which is what is the relevance criteria of the current query. It is interesting to note that in the case of true topical overlap, our approach acts implicitly in a retrieval-augmented setting, providing an example of how to complete a task and additional context with which to estimate relevance.

5.3 Extending Few-shot PRP to Point-wise and Set-wise Cases

Additionally, for the sake of completeness, we investigate the application of the few-shot approach on the other two modes of LLM inference for ranking, i.e., pointwise and setwise, as opposed to the pairwise mode reported so far. The results, as presented in Appendix B, show that none of these approaches benefit from the application of few-shot query-relevance examples most likely due to the complexity of these tasks itself as compared to the pairwise task - it is potentially easier to make a binary choice of preferring one document over the other as opposed to predicting a score (as in pointwise) or choosing a winner document from a set of more than 2 (usually of the order of 5 to 10) choices in case of setwise (similar to listwise).

6 Conclusions and Future work

We proposed a novel example selection process inspired by neural retrieval training processes, which improves unsupervised performance in a pair-wise ranking setting by exploiting in-context learning and is adaptable beyond a target domain. This non-parametric approach helps eliminate several decision choices involved in a supervised learning-to-rank pipeline, e.g., the architecture, the pre-training, index construction, negative sampling, distillation, etc. Despite the simplicity, our experiments confirm that the few-shot PRP not only significantly outperforms the zero-shot PRP on in-domain but also either statistically outperforms monoT5 (Covid dataset) or is statistically indistinguishable from it (SciFact dataset).

As future work, we plan to explore ways of selecting a variable number of examples on a per-query basis (Parry et al., 2024) or consider an open-domain ICL approach of using unlabelled data as contexts, (Long et al., 2023), e.g., information from Wikipedia, for improving the ranking task further.

Ethical Statement

Noting to declare.

Limitations

We mainly focus on the open-source lightweight LLMs ($\leq$7B) and whether the few-shot performance gains are much higher with larger LLMs (such as LLaMa-70B, GPT-3.5 or GPT-4) is yet to be investigated. We also consider only the 'All-Pairs' method for reranking the top-100 documents, which was one of the techniques used in (Qin et al., 2023). While (Qin et al., 2023) proposed a pseudo-sorting algorithm as an approximate strategy requiring with linear complexity (as opposed to quadratic complexity for an exhaustive pairwise setting) and (Zhuang et al., 2024a) proposed further improvements using more effective sorting algorithms, our approach can be trivially applied under these setting to improve efficiency.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Luiz Henrique Bonifacio, Hugo Queiroz Abonizio, Marzieh Fadaee, and Rodrigo Frassetto Nogueira. 2022. Inpars: Data augmentation for information retrieval using large language models. *CoRR*, abs/2202.05144.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, J Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners.

Advances in neural information processing systems, 33:1877–901.

Hyung Won Chung, Le Hou, Shayne 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, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

Nachshon Cohen, Hedda Cohen Indelman, Yaron Fairstein, and Guy Kushilevitz. 2024. Indi: Informative and diverse sampling for dense retrieval. In *ECIR (3)*, volume 14610 of *Lecture Notes in Computer Science*, pages 243–258. Springer.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the trec 2020 deep learning track. *Preprint*, arXiv:2102.07662.

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

Zhuyun Dai and Jamie Callan. 2019. Context-aware sentence/passage term importance estimation for first stage retrieval. *Preprint*, arXiv:1910.10687.

Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. 2023. Promptagator: Few-shot dense retrieval from 8 examples. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

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. 2024. Who determines what is relevant? humans or ai? why not both? *Commun. ACM*, 67(4):31–34.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. *SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking*, page 2288–2292. Association for Computing Machinery, New York, NY, USA.

Luyu Gao and Jamie Callan. 2021. Condenser: a pre-training architecture for dense retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 981–993, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Luyu Gao and Jamie Callan. 2022. Long document re-ranking with modular re-ranker. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2371–2376, New York, NY, USA. Association for Computing Machinery.

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. COIL: Revisit exact lexical match in information retrieval with contextualized inverted list. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3030–3042, Online. Association for Computational Linguistics.

Xingwei He, Yeyun Gong, A-Long Jin, Hang Zhang, Anlei Dong, Jian Jiao, Siu Ming Yiu, and Nan Duan. 2023. Capstone: Curriculum sampling for dense retrieval with document expansion. *Preprint*, arXiv:2212.09114.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 113–122, New York, NY, USA. Association for Computing Machinery.

Hidehito Honda, Toshihiko Matsuka, and Kazuhiro Ueda. 2017. Memory-based simple heuristics as attribute substitution: Competitive tests of binary choice inference models. *Cognitive Science*, 41(S5):1093–1118.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022a. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022b. Unsupervised dense information retrieval with contrastive learning. *Preprint*, arXiv:2112.09118.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

D. Kahneman and S. Frederick. 2002. Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, and D. Kahneman, editors, *Heuristics & Biases: The Psychology of Intuitive Judgment.*, pages 49–81. New York. Cambridge University Press.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020a. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in*

Natural Language Processing (EMNLP), pages 6769–6781, Online. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020b. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Carlos Lassance, Hervé Déjean, Stéphane Clinchant, and Nicola Tonellotto. 2024. Two-step SPLADE: simple, efficient and effective approximation of SPLADE. In *ECIR (2)*, volume 14609 of *Lecture Notes in Computer Science*, pages 349–363. Springer.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Minghan Li, Xueguang Ma, and Jimmy Lin. 2022. An encoder attribution analysis for dense passage retriever in open-domain question answering. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 1–11, Seattle, U.S.A. Association for Computational Linguistics.

Tianle Li, Xueguang Ma, Alex Zhuang, Yu Gu, Yu Su, and Wenhu Chen. 2023. Few-shot in-context learning on knowledge base question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6966–6980, Toronto, Canada. Association for Computational Linguistics.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2020. Distilling dense representations for ranking using tightly-coupled teachers. *Preprint*, arXiv:2010.11386.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *CoRR*, abs/2307.03172.

Quanyu Long, Wenya Wang, and Sinno Jialin Pan. 2023. Adapt in contexts: Retrieval-augmented domain adaptation via in-context learning. *arXiv preprint arXiv:2311.11551*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023a. Fine-tuning llama for multi-stage text retrieval. *arXiv preprint arXiv:2310.08319*.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023b. Zero-shot listwise document reranking with a large language model. *arXiv preprint arXiv:2305.02156*.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023c. Zero-shot listwise document reranking with a large language model. *CoRR*, abs/2305.02156.

Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder. 2020. Expansion via prediction of importance with contextualization. In *Proc. of SIGIR'20'*, pages 1573–1576.

Chuan Meng, Negar Arabzadeh, Arian Askari, Mohammad Aliannejadi, and Maarten de Rijke. 2024. Query performance prediction using relevance judgments generated by large language models. *CoRR*, abs/2404.01012.

Aristides Milios, Siva Reddy, and Dzmitry Bahdanau. 2023. In-context learning for text classification with many labels. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 173–184, Singapore. Association for Computational Linguistics.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2021. Large dual encoders are generalizable retrievers. *Preprint*, arXiv:2112.07899.

Feng Nie, Meixi Chen, Zhirui Zhang, and Xu Cheng. 2022. Improving few-shot performance of language models via nearest neighbor calibration. *Preprint*, arXiv:2212.02216.

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. *arXiv preprint arXiv:2003.06713*.

Rodrigo Frassetto Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *CoRR*, abs/1901.04085.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022a. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022b. Training language models to follow instructions with human feedback. In *Proc. of NIPS'22*.

Andrew Parry, Debasis Ganguly, and Manish Chandra. 2024. "in-context learning" or: How I learned to stop worrying and love "applied information retrieval". *CoRR*, abs/2405.01116.

Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. 2021. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *arXiv preprint arXiv:2101.05667*.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023a. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *Preprint*, arXiv:2309.15088.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023b. Rankzephyr: Effective and robust zero-shot listwise reranking is a breeze! *Preprint*, arXiv:2312.02724.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, et al. 2023. Large language models are effective text rankers with pairwise ranking prompting. *arXiv preprint arXiv:2306.17563*.

Mohammed Saeed, Giulio Alfarano, Khai Nguyen, Duc Pham, Raphael Troncy, and Paolo Papotti. 2021. Neural re-rankers for evidence retrieval in the FEVEROUS task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 108–112, Dominican Republic. Association for Computational Linguistics.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agent. *arXiv preprint arXiv:2304.09542*.

Yuting Tang, Ratish Puduppully, Zhengyuan Liu, and Nancy Chen. 2023. In-context learning of large language models for controlled dialogue summarization: A holistic benchmark and empirical analysis. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 56–67, Singapore. Association for Computational Linguistics.

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)*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem 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, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *Preprint*, arXiv:2310.16944.

Christophe Van Gysel and Maarten de Rijke. 2018. Pytrec_eval: An extremely fast python interface to trec_eval. In *SIGIR*. ACM.

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.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2023. SimLM: Pre-training with representation bottleneck for dense passage retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2244–2258, Toronto, Canada. Association for Computational Linguistics.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, K. Funk,

Rodney Michael Kinney, Ziyang Liu, W. Merrill, P. Mooney, D. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, B. Stilson, A. Wade, K. Wang, Christopher Wilhelm, Boya Xie, D. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. Cord-19: The covid-19 open research dataset. *ArXiv*.

Shitao Xiao, Zheng Liu, Weihao Han, Jianjin Zhang, Defu Lian, Yeyun Gong, Qi Chen, Fan Yang, Hao Sun, Yingxia Shao, and Xing Xie. 2022. Distill-vq: Learning retrieval oriented vector quantization by distilling knowledge from dense embeddings. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 1513–1523, New York, NY, USA. Association for Computing Machinery.

Tingyu Xie, Qi Li, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Self-improving for zero-shot named entity recognition with large language models. *CoRR*, abs/2311.08921.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *Preprint*, arXiv:2007.00808.

Oleg Zendel, Anna Shtok, Fiana Raiber, Oren Kurland, and J. Shane Culpepper. 2019. Information needs, queries, and query performance prediction. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 395–404, New York, NY, USA. Association for Computing Machinery.

Hansi Zeng, Hamed Zamani, and Vishwa Vinay. 2022. Curriculum learning for dense retrieval distillation. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 1979–1983. ACM.

Xinyu Zhang, Sebastian Hofstätter, Patrick Lewis, Raphael Tang, and Jimmy Lin. 2023. Rank-without-gpt: Building gpt-independent listwise rerankers on open-source large language models. *CoRR*, abs/2312.02969.

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2022. Rankt5: Fine-tuning t5 for text ranking with ranking losses. *Preprint*, arXiv:2210.10634.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. 2024a. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. 2024b. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024. ACM.

A Implementation Details

We apply PyTerrier bindings over each neural model and use Terrier's 'pytrec_eval' (Van Gysel and de Rijke, 2018) extension for computing the evaluation metrics. We used the HuggingFace implementations of monoT5 (Nogueira et al., 2020), Zephyr (Tunstall et al., 2023) and Flan-T5-XL (Chung et al., 2022). All models are executed on a single RTX 4090 GPU.

B Pointwise and Setwise Few-shot

A relative preference between a pair of documents is an easier decision choice than estimating the relevance of a document to a query, making pairwise ranking a natural choice. However, observing the effect of few-shot ICL examples in the pointwise and listwise methods is necessary. Experiments using both pointwise and listwise - specifically, the approach proposed in (Zhuang et al., 2024b) - are carried out to support our claims further.

As per our experiment, we observe that FLAN-T5's point-wise estimation is not good, and due to its short max-context length, it is also ineffective for the list-wise setting. We now include these findings over TREC Deep Learning with only Zephyr in point and list-wise settings.

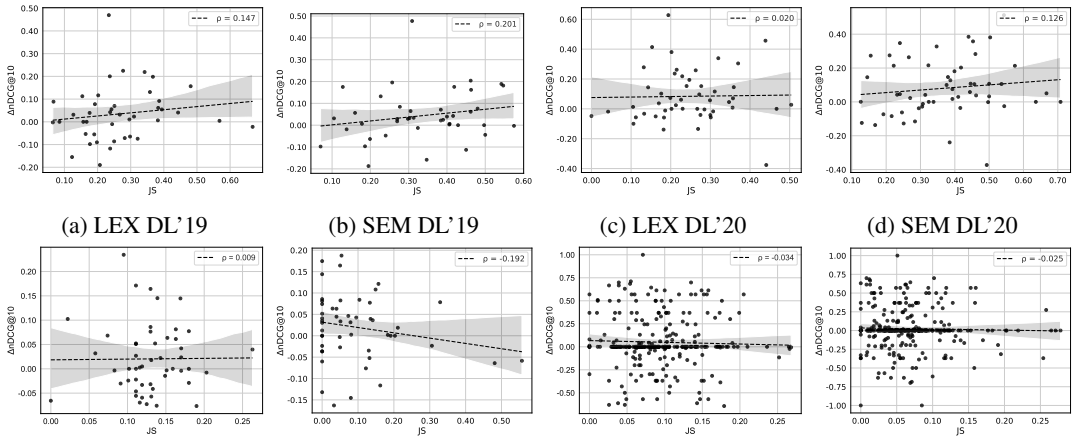
Table 5 shows that additional examples provide small improvements over a single example, with the only exception being DL'19 in the pointwise setting where we see an actual improvement. However, performance in pointwise and listwise is significantly reduced over pairwise in a zero-shot or few-shot setting. We particularly observe improvements in nDCG@10 using localized examples compared to a random 1-shot example. We see that AP@100 is unstable under different example settings, coupled with significantly reduced effectiveness compared to a pair-wise method, and we argue that the pairwise method turns out to be the most effective alternative in an ICL setting.

C Per-query Analysis

To understand how the locality of examples to the original query correlates with retrieval perfor-

Table 5: Analysis of our few-shot extension to Pointwise and Setwise ranking strategies.

Type	Retriever	TREC DL'19 $\bar{J}(\mathcal{N}_k)$	TREC DL'19 AP@100	TREC DL'19 nDCG@10	TREC DL'20 $\bar{J}(\mathcal{N}_k)$	TREC DL'20 AP@100	TREC DL'20 nDCG@10
Pointwise	Zephyr-0S	n/a	.2268	.5195	n/a	.2008	.5690
	Zephyr-1S	.041	.2402	.5271	.370	.1893	.5503
	Zephyr-LEX-1S	.267	.2416	.5559	.244	.1841	.5610
Setwise	Zephyr-0S	n/a	.3122	.6143	n/a	.3468	.5953
	Zephyr-1S	.041	.2292	.6176	.029	.3517	.6067
	Zephyr-LEX-1S	.267	.3080	.6417	.244	.3613	.6101
	Zephyr-SEM-1S	.352	.2993	.6317	.370	.3373	.5985



(a) LEX DL'19 (b) SEM DL'19 (c) LEX DL'20 (d) SEM DL'20

(e) LEX Covid (f) SEM Covid (g) LEX SciFact (h) SEM SciFact

Figure 4: Per-query analysis showing the relation between Jaccard similarity (JS) of current query and 1-shot example with the ΔnDCG@10 relative to 0S with Zephyr-1S using the semantic and lexical neighborhoods for in-domain and out-domain test sets. The Pearson correlation (ρ) is shown in each case.

mance, we measured ΔnDCG@10 per-query basis between the 0S and LEX/SEM-1S with JS between the current query and examples as seen in Figure 4. Our observation indicates that localized examples with higher JS scores yield better performance gains per query as well as for entire topics as observed from $\bar{J}(\mathcal{N}_k)$ scores in Table 2 and 3. We observe a positive correlation in-domain and a minor correlation in the OOD setup, suggesting that we find better examples if we have a domain-specific training set. Interestingly, we observe a gain in nDCG@10 scores over a greater fraction of the queries, improving our overall retrieval performance in all the experiments.

D Few-shot PRP Effect on a Dense First-stage Retriever

Our main results were reported with BM25 being the first stage retriever. In this section, we supplement our main results with findings on replacing BM25 with an unsupervised dense ranker, namely the Contriever model (Izacard et al., 2022b). The objective is to find out if our proposed few-shot reranking methodology also works effectively another ranker with different characteristics.

Table 6 reports that applying few-shot pairwise prompting yields consistent improvements even on the candidate set of top-documents retrieved with a dense retrieval model (the table also includes the BM25 + Few-shot PRP results from Table 2 for the sake of completeness). Contriever yields a higher retrieval effectiveness than BM25 to start with, and applying 1-shot prompting with a semantic neighborhood on Zephyr leads to the largest improvement (nDCG@10 of 0.6889). This result is considerably close to that of monoT5, which is a supervised model. The results in Table 6 further substantiates our claim that in-context learning based reranking can achieve comparable results to supervised approaches without requiring any parametric training.

While we observe that replacing BM25 with a dense ranker, such as Contriever, provides further improvements bringing the effectiveness closer to

Table 6: A comparison to show the effect of changing the phase-one retriever to Contriver with two different neighborhood similarity functions. Each one-shot result reported in this table is an average over 5 runs with the standard deviations included in superscript. The best scores across all unsupervised approaches are bold-faced, and the overall best result in a group is underlined. Letters a to d are used to indicate the statistical significance of a retriever with Zephyr-0S,Zephyr-LEX-1S, Zephyr-SEM-1S, and monoT5.

		TREC DL'19			TREC DL'20		
Type	Retriever	$\bar{J}(\mathcal{N}_k)$	AP@100	nDCG@10	$\bar{J}(\mathcal{N}_k)$	AP@100	nDCG@10
Baseline	Contriever	n/a	.3400	.5888	n/a	.3694	.5845
	Zephyr-0S	n/a	.3693	.6391	n/a	.3637	.5758
Ours	BM25 » Few-shot PRP						
	Zephyr-LEX-1S	.267	$.3447^{(.0019)\,a}$	$.6742^{(.0005)\,a}$	.244	$.3793^{(.0052)\,abc}$	$.6457^{(.0077)\,abc}$
	Zephyr-SEM-1S	.352	$\underline{.3512}^{(.0041)\,a}$	$\underline{.6785}^{(.0028)\,a}$	.370	$\underline{.3824}^{(.0019)\,abc}$	$\underline{.6480}^{(.0033)\,abc}$
	Contriever » Few-shot PRP						
	Zephyr-LEX-1S	.267	$\mathbf{.4145}^{(.0013)\,abcd}$	$\mathbf{.6889}^{(.0021)\,a}$	.244	$.4219^{(.0015)\,abc}$	$.6677^{(.0018)\,abc}$
	Zephyr-SEM-1S	.352	$.3940^{(.0012)\,ab}$	$.6679^{(.0036)\,a}$	.370	$\mathbf{.4231}^{(.0016)\,abc}$	$\mathbf{.6680}^{(.0010)\,abc}$
Supervised	monoT5	n/a	$\mathbf{.3570}^{a}$	$\mathbf{.6998}^{a}$	n/a	$\mathbf{.3970}^{a}$	$\mathbf{.6729}^{a}$

Table 7: Evaluating (nDCG@10) re-ranking performance on top-20 Contriever retrieved documents in out-of-domain settings. Letters a to d are used to indicate the statistical significance of a retriever with Zephyr-0S, Zephyr-LEX-1S, Zephyr-SEM-1S, and monoT5 (paired t-test with $p = 0.05$).

	TREC Covid		SciFact	
Retriever	$\bar{J}(\mathcal{N}_k)$	nDCG@10	$\bar{J}(\mathcal{N}_k)$	nDCG@10
BM25	n/a	.5781	n/a	.6722
Contriever	n/a	.3723	n/a	.6081
BM25 » Few-shot PRP				
Zephyr-0S	n/a	.6571	n/a	.6872
Zephyr-LEX-1S	.130	$\mathbf{.6790}^{ad}$	.093	**.6988**
Zephyr-SEM-1S	.094	$.6753^{d}$	.067	.6880
Contriever » Few-shot PRP				
Zephyr-0S	n/a	.4989	n/a	.5628
Zephyr-LEX-1S	.130	.4963	.093	.6264
Zephyr-SEM-1S	.094	.4922	.067	.6027
monoT5	n/a	.6376	n/a	$\mathbf{.7204}^{ac}$

monoT5 for in-domain evaluation, the out-domain effectiveness of Contriever + Few-shot PRP is lower than that of BM25 + Few-shot PRP (see Table 7). The main reason for this is the lower performance of Contriever on OOD collections, e.g., 0.3723 with Contriever vs. 0.5781 with BM25 on TREC Covid. Despite the retrieval effectiveness improving due to reranking, the overall results are still lower as compared to the BM25 >> Few-shot PRP pipeline.

Additionally, similar to our observations for BM25, even for Contriever we find that examples selected based on lexical similarity leads to more consistent and robust behaviour across domains. In contrast, examples selected by semantic similarity exhibit larger variations in performance across domains.