DiSCo: LLM Knowledge Distillation for Efficient Sparse Retrieval in Conversational Search

Simon Lupart
University of Amsterdam
Amsterdam, Netherlands
s.c.lupart@uva.nl

Mohammad Aliannejadi
University of Amsterdam
Amsterdam, Netherlands
m.aliannejadi@uva.nl

Evangelos Kanoulas
University of Amsterdam
Amsterdam, Netherlands
e.kanoulas@uva.nl

Abstract

Conversational Search (CS) involves retrieving relevant documents from a corpus while considering the conversational context, integrating retrieval with context modeling. Recent advancements in Large Language Models (LLMs) have significantly enhanced CS by enabling query rewriting based on conversational context. However, employing LLMs during inference poses efficiency challenges. Existing solutions mitigate this issue by distilling embeddings derived from human-rewritten queries, focusing primarily on learning the context modeling task. These methods, however, often separate the contrastive retrieval task from the distillation process, treating it as an independent loss term. To overcome these limitations, we introduce DiSCo (Distillation of Sparse Conversational retrieval), a novel approach that unifies retrieval and context modeling through a relaxed distillation objective. Instead of relying exclusively on representation learning, our method distills similarity scores between conversations and documents, providing more freedom in the representation space and better leveraging the contrastive nature of document relevance. Extensive experiments on Learned Sparse Retrieval (LSR) across five CS datasets demonstrate that DiSCo achieves substantial improvements in both in-domain and out-of-domain retrieval tasks, achieving up to a six-point gain in recall for out-of-domain datasets over state-of-the-art methods. Additionally, DiSCo employs a multi-teacher distillation strategy, using multiple LLMs as teachers, further enhancing performance and surpassing the individual teachers in in-domain settings. Furthermore, analysis of model sparsity reveals that DiSCo allows for more effective control over the sparsity of the trained models.

CCS Concepts

• **Information systems** → **Query representation**; **Language models**; **Information retrieval**.

Keywords

conversational search, query understanding, neural sparse retrieval

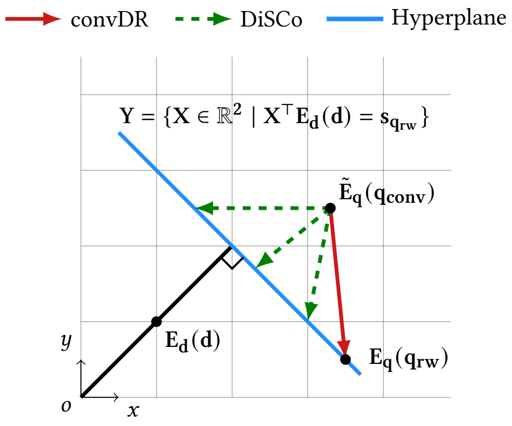

Figure 1: Similarity Score Distillation in $\mathbb{R}^2$. Existing loss functions bound representation of the full conversation representation to converge to a single rewrite representation (convDR, red arrow), while if we consider document d as anchor, an infinite number of representations, other than $E_q(q_{rw})$, have the same similarity with d (Y, blue hyperplane). DiSCo allows the model to converge to the best representation from the Y hyperplane (green arrows), as a relaxation.

ACM Reference Format:
Simon Lupart, Mohammad Aliannejadi, and Evangelos Kanoulas. 2025. DiSCo: LLM Knowledge Distillation for Efficient Sparse Retrieval in Conversational Search. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3726302.3729966

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3729966

1 Introduction

Conversational Search (CS) is a well-established task, that has seen major improvements recently, thanks to the development of Large Language Models (LLMs) [3, 41, 43]. The goal of the CS task is to retrieve relevant documents from a corpus within a conversational context, in response to the user's latest utterance. While sharing similarities with ad-hoc retrieval, the main challenge of CS remains to model the conversational context [8, 54, 64]. More specifically, as the conversation advances, the context becomes longer and noisier, often with topic switches and language ambiguities, making the last user utterance complex to resolve and understand for retrieval systems [2, 10, 18].

Training retrieval models for this task is, however, challenging, because of the lack of large-scale conversational datasets, making human annotation an essential component of the process. In particular, a lot of effort has been invested first to create conversations with passage relevance judgment [3, 10], and then to rewrite for each user utterance the contextualized version of each query (rewritten query) as the optimal denoised query of each turn [14]. Datasets with rewrites enable us to learn the Conversational Query

Rewriting (CQR) task by auto-regressive models (e.g., T5 [55]), and then only pass the rewritten queries to retrieval models instead of the full noisy conversations. However, this two-step approach – rewrite and retrieve – is not efficient [65], and may lead to information loss and error within the rewrite phase that can propagate to the retrieval phase. Hence, the need for unified retrieval models that do both tasks together in the representation space [46, 52].

Approaches such as ConvDR [65], coSPLADE [19] and LeCoRe [42] all learn both conversational context modeling and retrieval tasks within the representation space, either in a dense or sparse embedding space. As illustrated in Figure 1, they use a distillation objective enforcing the conversation representations to converge to the representations of gold human-rewritten queries (red arrow in the figure). This objective is, however, restrictive and assumes gold-rewritten query representations are the *only* optimal rewriting and *only* learning target in the representations space, leaving little freedom for the model to further learn and optimize the representations. DiSCo, differently from ConvDR relaxes the distillation targeting an entire hyperplane (blue hyperplane in the figure) instead of a single representation. This is achieved by focusing on query document similarity scores, rather than solely on the rewrite. Besides, previous work [19, 42, 65] distills query rewrite independently from the ranking objective. We thus propose to fill this gap, with a single distillation loss that unifies both context modeling and ranking tasks, while not limiting the model to learn one single representation.

Our method learns to distill the similarity scores between rewritten queries and documents rather than query representations directly. By distilling similarities, we first align with the contrastive nature of the ranking task [36, 57, 58], but we also relax the training objective toward the final goal of retrieval, which is to compute similarities between queries and documents [24, 56, 59]. This relaxation allows the student to further learn the context modeling task, considering relevant and irrelevant documents from the corpus, rather than mapping all representations to the human representations. This objective also follows the precepts from Hofstätter et al. [20] on distillation for ad-hoc retrieval and can benefit from hard negatives within the distillation loss. Both dense and sparse methods could be subject to this relaxation; however, in this work, we focus on learned sparse architectures [66], such as SPLADE [15, 31], as the relaxation would have more degrees of freedom due to the high dimensionality of the representations for LSR.

Besides, the proposed relaxation also reduces the level of constraints on the teachers, as any method producing similarity scores could be used for the distillation. In the context of CS, this allows the model to learn from multiple LLM rewrites, by fusing similarity scores of several teachers into a single score to be distilled. To the best of our knowledge, our work is the first to distill the knowledge of multiple LLM teachers for query rewriting. This also distinguishes our method from the work from Hofstätter et al. [20], as we distill from LLM knowledge through the rewrites.

Through our work, we make the following contributions:

- We propose DiSCo, a relaxation of the training objective of CS model distilling similarities of rewritten queries rather than representations. [1]
- We propose to distill knowledge from multiple teachers, unlocking the potential of mixtures of LLMs in CS.
- We evaluate the effectiveness of DiSCo on in-domain (QReCC, TopiOCQA) and out-of-domain (CAsT 2020, 2022, and iKAT 2023) datasets, achieving state-of-the-art performance.
- We analyze the sparsity of the learned representations compared to those of the original teacher models, and demonstrate the efficiency gains of the approach.

Our results demonstrate that our **Di**stillation of **S**parse **Co**nversational retrieval, **DiSCo**, leads to in-domain performance improvements, through our relaxation of the distillation loss. We also see improved generalization capacities with 12% gains on recall and 16% on precision compared to previous zero-shot models (LeCoRe and QRACDR) on CAsT 2020. We demonstrate that distilling from multiple LLM teachers for the query rewriting task brings further performance gains, compared to single LLM teacher. Finally, thanks to the increased freedom on the representations, we show that we can better control the model sparsity, even for long contexts, in deeper turns of the conversations.

2 Related Work

Conversational Search (CS) differentiates itself from ad-hoc search primarily through the nature of the input. Conversational history can grow very long with turns' dependencies, whereas search engine queries are typically concise, limited to a few words [6, 39]. The main challenge in addressing CS is modeling the conversational context, to only represent useful information from the previous turns [61]. Two possibilities exist to learn this noise reduction: first on the token level, with generative models that learn to generate contextualized queries from past conversations [14], or within the representation spaces [65] of retrieval models [44].

Query rewriting. Learning to resolve the conversational history is difficult, as no or very few conversational search engines are in production today, limiting the availability of user-interaction data. The effort is thus mostly based on human-annotated data. One of the first conversational datasets are QuAC [7] and ORConvQA [53], where human annotators were tasked to create conversations out of existing documents, assuming a human-machine conversation over the topic of the document. Each paragraph could potentially be an utterance and relevant passages would be inferred by construction. Later CANARD [14] was released on top of ORConvQA, consisting of human rewrites of each utterance of the conversations, to help learn the context modeling task. This defined one of the main sub-tasks of CS: Conversational Query Rewriting (CQR) [14, 64]. Following studies learned the CQR task at the token level – generating automatic query rewrites based on the conversation history – on autoregressive language models (e.g., T5 or BART) [40, 61]. Similar methods [45] used language models to answer the query, before resolving the context. Today, work is still being done to solve the CQR task, using LLMs in zero- or few-shot fashion. CHIQ [43] proposes to decompose the context modeling task into several simpler sub-tasks for the LLMs – history enhancement, answer generation, question disambiguation – to gain in interpretability and effectiveness, LLM4CS aggregates several rewrites and answers within the representation space of the retrieval models [41], and

[1] Models checkpoints at https://huggingface.co/slupart/splade-disco-human-mistral

MQ4CS focuses on multi-aspect query rewrites [1]. However, context modeling using LLMs is too computationally costly, making it impossible for production.

Distillation in CS. Lots of research aims at modeling the conversational context via learning to represent the conversation based on gold rewritten queries. Existing methods achieve this goal by taking the representation of the rewritten gold query as a teacher model and learning to distill the representation [19, 42, 46, 65]. ConvDR [65] learns the mapping between full conversation and human rewrites representations by minimizing an Mean Square Error (MSE) loss on the CLS tokens of both, performing well for dense bi-encoders. QRACDR [46], also proposes a similar distillation, with several new MSE terms, between documents and queries to improve the representations and better align with the contrastive nature of the task. We differentiate from these works as we inspire from contrastive margin distillation (Margin-MSE [20]) and relax the query rewriting distillation by learning based on the relevance scores (i.e., the dot product of the query representation and a document), rather than the representation itself. This way, not only do we enable converging towards indefinite possible query representations, but also we take advantage of the contrastive nature of ranking by aligning the rewrite task to retrieval. Furthermore, we are not bound to one query or one teacher.

Learned sparse representation. As Learned Sparse Retrieval (LSR) gained popularity in ad-hoc retrieval [28, 29, 31], SPLADE architectures were proposed for conversational passage retrieval [19, 37, 42], to benefit their interpretability and robustness properties [16, 17, 38]. In coSPLADE [19], the authors use an MSE loss between full conversation and human rewrite representations of the sparse bag-of-words representations (of dimension 30k), and show promising performance. In the meantime, LeCoRe [42] also aims to distill human rewrites on sparse representations, through intermediate embedding layers and by filtering a maximum of dimensions, to avoid the MSE on large dimensionality vectors. Distilling sparse representations is, however, challenging, as it requires controlling which dimensions should be activated, often restricting the potential of the sparse representations. Additionally, using a MSE is problematic since it treats each dimension separately, neglecting the significance of the associated dimension. These issues limit the effectiveness of sparse retrieval in previous work. In our work, by only distilling end scores, we give the model complete freedom to learn which dimensions to activate. We also investigate sparsity further, including methods to control it within CS.

3 Proposed Method

In this section, we first recall notations from the CS and LSR fields, before presenting DiSCo and our relaxed distillation objective.

3.1 Preliminaries

Notation. We consider a set of conversations between a user and a system, each composed of multiple turns. At turn n, we have access to previous queries and answers, together with the final user utterance q_n. We denote the full conversational context as q_{conv}, separated with [SEP] tokens, and q_{rw} as the gold rewritten query. q_{rw} resolves various complexities such as ellipsis and ambiguities of the conversation and can be generated by either human or LLM.

$$q_{conv} = q_n, a_{n-1}, q_{n-1}, ..., a_0, q_0 \ .$$

Furthermore, similarly to other approaches in CS [19, 42, 46, 65] we rely on two already trained encoder models E_q and E_d, for queries and documents representations. These backbone models are trained on a large-scale IR dataset with regular short-form queries extracted from search engine logs. We also follow the assumption made in CS that E_d is already good at representing documents and doesn't need further fine-tuning, only E_q needs further fine-tuning to adapt to long conversational contexts, noted $\tilde{E}_q$. This can be done with a contrastive InfoNCE [48] loss on the CS datasets (e.g., convANCE, convSPLADE), or with a distillation loss (e.g., ConvDR and other distillation approaches such as DiSCo).

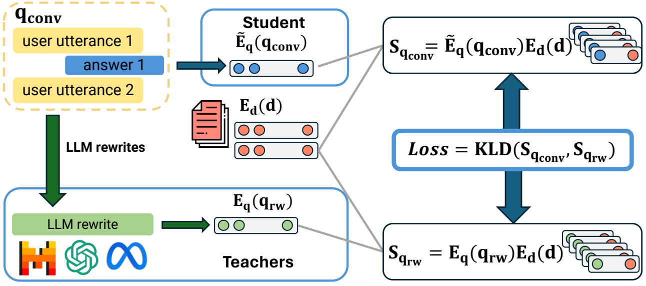
$\mathbf{q_{conv}}$
user utterance 1
answer 1
user utterance 2
Student
$\tilde{\mathbf{E}}_\mathbf{q}(\mathbf{q_{conv}})$
$\mathbf{E_d(d)}$
LLM rewrites
$\mathbf{E_q(q_{rw})}$
LLM rewrite
Teachers
$\mathbf{S_{q_{conv}} = \tilde{E}_q(q_{conv})E_d(d)}$
$\boldsymbol{Loss} = \mathbf{KLD(S_{q_{conv}}, S_{q_{rw}})}$
$\mathbf{S_{q_{rw}} = E_q(q_{rw})E_d(d)}$

Figure 2: DiSCo, as the Distillation of LLM rewritten queries through a contrastive objective. Previous works distilled representations themselves, while our approach distills similarities with documents from the corpus, relaxing the learning objective.

Learned Sparse Retrieval. We rely on the SPLADE architecture [15, 17], based on the BERT pretrained transformers [12, 34]. SPLADE makes use of the Masked Language Modeling embedding layer of BERT to create sparse representations, where each dimension corresponds to a token of the vocabulary. E_q, E_d and $\tilde{E}_q$ are the MLM outputs of the model, of dimension 30k (vocabulary dimension). To control the sparsity of the model, the authors propose to use a regularization loss [50], The sparsity of the models can be controlled with the two hyperparameters λ_q and λ_d, as weights for the regularization loss. As we rely on the same SPLADE model, we also included this loss.

3.2 DiSCo

We describe our distillation process in Figure 2. The conversation is passed through the teacher model, rewriting the last user utterance, and computing similarities with several documents from the corpus. Meanwhile, the student encodes the full conversation and scores it with those documents. The final training objective for the student is to match the teacher similarity scores. This relaxes the previous objective, which distilled representations directly.

Representation distillation. Existing approaches [19, 42, 46, 65] distill query rewrites on the representation space, with a direct constraint on the representation. The goal is to have the representations of the full conversations converging toward the representations of

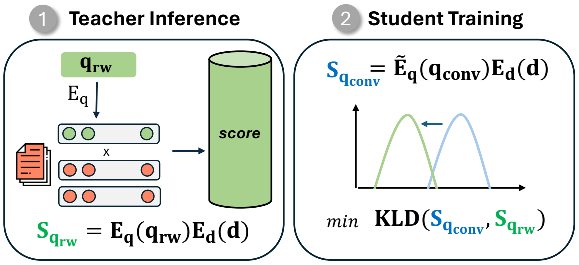
1 Teacher Inference
2 Student Training
q_{rw}
E_q
x
score
$S_{q_{rw}} = E_q(q_{rw})E_d(d)$
$S_{q_{conv}} = \tilde{E}_q(q_{conv})E_d(d)$
min KLD($S_{q_{conv}}$, $S_{q_{rw}}$)

Figure 3: Distillation process. The first step stores scores from the rewritten queries with documents from the corpus. Then the student query encoder $\tilde{E_q}$ is trained to reproduce the output scores of the teacher.

the rewritten queries:

$$\tilde{E_q}(q_{conv}) \longrightarrow E_q(q_{rw}) \, . \tag{1}$$

Relaxation of the distillation. In our work, we propose to distill the scores rather than the representations. This allows for a relaxation of the training objective. Instead of forcing $\tilde{E_q}(q_{conv})$ to converge to $E_q(q_{rw})$, we allow the entire hyperplane on which the similarity for q_{conv} and q_{rw} with an anchor document d is equal:

$$\tilde{E_q}(q_{conv}) \longrightarrow \{X \in \mathbb{R}^h \mid X^\top E_d(d) = s_{q_{rw}}\} \, , \tag{2}$$

with targeted scores $s_{q_{rw}} = E_q(q_{rw})^\top E_d(d)$, and h the dimension of the representation space. More intuitively, while the existing distillation loss functions bound the model to converge to one single embedding [65], our loss allows for infinite possible optimum embeddings, as long as they have the same dot product with the target relevant document embedding (i.e., any point in the hyperplane). Such distillation on scores can be trained with a Kullback Leibler Divergence loss [25] on the distribution of scores:

$$\mathcal{L}_{\text{KLD}} = D_{\text{KL}}(\mathcal{S}_{q_{rw}} \,||\, \mathcal{S}_{q_{conv}}) \, , \tag{3}$$

where $\mathcal{S}_{q_{rw}}$ and $\mathcal{S}_{q_{conv}}$ are the distributions of similarity scores within the batch. It differs from the existing learning objective which is achieved with an MSE loss on each dimension of the vector representations independently [19, 65]:

$$\mathcal{L}_{\text{MSE}} = \text{MSE}(\, \tilde{E_q}(q_{conv}), \, E_q(q_{rw})) \, . \tag{4}$$

This new $\mathcal{L}_{\text{KLD}}$ loss includes the benefit of the contrastive objective, anchoring itself on relevant and hard documents from the corpus. Also note that in the previous optimization, the final loss was the sum of the distillation MSE loss with the contrastive InfoNCE loss, while we only use a single contrastive distillation KLD loss, unifying both objectives into one distillation loss. Figure 3 illustrates the distillation process. First, the teacher retrieves documents using the rewritten queries and stores the similarity scores. Then, the student model encodes the full conversation, learning to have equal similarities to the teacher on these documents, by minimizing the KLD distillation loss. This defines the final learning objective of our DiSCo models, as minimizing the distributions of scores between the student and teacher models.

Hard negatives. As our new training objective from Equation 3 is now contrastive, we can benefit from hard negative mining. This was not possible with the previous distillation objective, as it only uses conversation representations, independently from documents.

Thus applied to the case of multiple negatives per query, the target space from Equation 2 becomes a subspace of **(h-N)** dimensions. This still keeps a higher degree of freedom, since the number of negatives is much lower than the dimension of the embedding space ($h >> N$).

$$\{X \in \mathbb{R}^h \mid X^\top E_d(d_i) = s^i_{q_{rw}} \;\forall i \in [1, N]\} \, , \tag{5}$$

with $s^i_{q_{rw}} = E_q(q_{rw})^\top E_d(d_i)$ being the scores from multiple hard negatives. These negatives are mined during the Teacher Inference step from Figure 3, further improving training objective [16, 63].

Teacher Models. An important component of distillation relies on the choice of the teacher. The teacher is first used to rewrite the conversation utterance, into q_{rw}, which will be encoded $E_q(q_{rw})$ and distilled through the similarities scores $s_{q_{rw}}$. We use multiple LLMs, together with human rewrite (when available), as teachers. Having several teachers is motivated by the well-established effectiveness of ensembling strategies, where stronger teachers would provide higher-quality training signals, leading to better student model performance. This results in several trained DiSCo students, depending on the teacher used. Below we list all the teachers:

- **T0**: T5QR
- **T1**: LlamaQR
- **T2**: MistralQR
- **T3**: HumanQR

Furthermore, another property of the proposed distillation is that we can distill from multiple teachers, making the distill from several LLMs or humans possible. We do so by averaging the similarity scores of the T teachers, each having a different rewrite, thus a different similarity with the documents. We experimented with mean, min, and max aggregation methods, observing no significant differences in performance, and thus decided to use the more simple mean aggregation. Overall, the distilled score is:

$$s_{q_{rw_{all}}} = \frac{s_{q_{\text{rw}_1}} + s_{q_{\text{rw}_2}} + \cdots + s_{q_{\text{rw}_T}}}{T} \, . \tag{6}$$

The final learning objective of our DiSCo multi-teacher is the same as for the regular DiSCo (i.e., KLD from Equation 3), but distilling the average score of the set of teachers, instead of using a single similarity from a unique teacher.

Student-Teacher Fusion. Finally, we propose a fusion of the teacher and the student model, combining the ranked lists of both at inference. This is achieved using an average normalized score over the ranked list of both models.

4 Experiments Design

4.1 Datasets and Metrics

Datasets. We evaluate our method's effectiveness on various conversational passage retrieval datasets, detailed in Table 1, namely:

- **QReCC** [4] is built upon ORConvQA [53], NQ [26] and TREC CAsT 2019 [10]. It features 13K conversations and 80K turns created by human annotators around existing documents.
- **TopiOCQA** [2] distinguishes itself from QReCC focusing on topic switches, where every Wikipedia hyperlink is considered a

Table 1: Statistics of the datasets.

Dataset	Split	# Conv.	# Turns	Collection
QReCC	Train	10,823	63,501	54M
	Test	2,775	16,451	
TopiOCQA	Train	3,509	45,450	25M
	Test	205	2,514	
TREC CAsT 20	Test	25	216	38M
TREC CAsT 22	Test	18	205	138M
TREC iKAT 23	Test	25	176	116M

topic switch in a conversation built around that Wikipedia topic. It features 4K conversations and 48K turns.

- **TREC CAsT 2020, 2022** [9, 49] and **TREC iKAT 2023** [3] are smaller conversational datasets, but carefully hand-crafted to include various conversational complexities. They are coupled with high-quality relevance judgements done by NIST assessors.

Effectiveness Metrics. We report the results in terms of the main IR metrics [42, 65], as well as official TREC CAsT and iKAT ones [3, 10]. Metrics are Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain (nDCG) [22] at 3. As we focus on first-stage retrieval, we also include Recall at different ranks [62]. We determine the statistically significant differences by doing a two-sided paired t-test with Bonferroni correction at 95% confidence ($p < 0.05$).

Efficiency Metrics. In terms of inference efficiency, we report the number of *LLM inference calls* when used for query rewriting, as it is a strong indicator of inference latency. As an example, it takes an average of 4.4 seconds to generate 64 tokens on Llama 3.1, on A100 GPU, while the typical dual-encoder retrieval latency is in the range of 100 milliseconds on CPU [27, 30, 32].

We also provide Rewrite and Retrieval efficiency, of several models. This is performed per query, averaged over the entire test set. Rewrite is the latency for rewriting, while retrieval is both query encoding and search through the inverted index (in the case of SPLADE models). We used a 4th AMD EPYC CPU and an A100 GPU.

We also report the FLOPs [15, 50] values, a metric for efficiency in learned sparse retrieval, which intuitively estimates the number of floating point operations between a query and a document in an inverted index. FLOPs can be computed as follows:

$$FLOPs = \mathbb{E}_{q,d}\left[\sum_{j \in V} p_j^{(q)} p_j^{(d)}\right] , \qquad (7)$$

where $p_j^{(q)}$ and $p_j^{(d)}$ are the probabilities of activation of the j^{th} token in the vocabulary, resp. in query and document representations, over which we average on the dataset distribution.

4.2 Baselines

We compare our DiSCo model with a wide range of competitive methods, consisting of query rewriting, supervised fine-tuned, and distillation-based methods, which we list below.

Table 2: One-shot prompt used for rewriting with the LLMs teachers. The example is taken out of the QReCC dataset.

```
# Instruction: I will give you a conversation between a
user and a system. You should rewrite the last question
of the user into a self-contained query.
# Example 1:
# Context:
user: Tell me about the benefit of Yoga?
system: Increased flexibility, muscle strength.
# Please rewrite the following user question:
Does it help in reducing stress?
# Re-written query:
Does Yoga help in reducing stress?
# Example 2:
# Context:
<ctx>
# Please rewrite the following user question:
<utterance>
# Re-written query:
```

Query rewriting methods. *SPLADE-[T5/Llama/Mistral/Human]QR* does retrieval using the SPLADE ad-hoc retrieval model trained on MS MARCO [47]. As input, we pass the rewritten query using T5 [55], Llama 3.1 [13], Mistral [23], or gold human rewrites[2]. *SPLADE no rewrite* is the same ad-hoc retrieval model without rewrite, on the original conversation. *IterCQR* [21], *CHIQ-Fusion* [43], and *LLM4CS* [41] are state-of-the-art query rewriting baselines; however, they require multiple LLM calls at inference, which puts them under a high disadvantage. For *LLM4CS*, we reproduced their best setting (RAR, Mean aggregation from $N = 5$, CoT, GPT-4).

Supervised fine-tuned methods. *convSPLADE* [42] and *convANCE* [42] are two methods fine-tuned using the InfoNCE loss on the conversational contrastive labels. Their input is the whole conversational context. We reproduce convSPLADE for out-of-domain retrieval using the same hyperparameters as in the original paper.

Distillation-based method. *LeCoRe* [42], *QRACDR* [46] and *ConvDR* [65][3] all learn to distill the gold human rewrite representations and usually combine the distillation loss with an InfoNCE. Like our DiSCo, these models do not rely on LLM calls either, making them comparable. *ConvDR* and *QRACDR* are both dense approaches, using latent representations, while *LeCoRe* uses learned sparse representations. All of them distill query representations, while DiSCo the similarity scores, as a relaxation of their distillation method.

4.3 Implementations details

We use the SPLADE++ [16][4] checkpoints from Huggingface [60] for all our SPLADE models. To further finetune on the CS task, we fine-tune for 5 epochs, with a learning rate of $2e^{-5}$, a max sequence length for queries of 64 tokens, and 100 tokens for answers, with

[2] We used the latest Llama `meta-llama/Meta-Llama-3.1-8B-Instruct`, Mistral `mistralai/Mistral-7B-Instruct-v0.2` and T5 `castorini/t5-base-canard`

[3] We did not manage to reproduce the results reported in the original papers, as our reproduced results were much lower. Therefore, we report the numbers in the original papers. Therefore, we are not able to perform any significance tests with these models.

[4] `naver/splade-cocondenser-ensembledistil` [16]

Table 3: In-Domain Performance on TopiOCQA and QReCC. DiSCo multi-teach for TopiOCQA is the combination of T_1 and T_2, and QReCC uses T_2 and T_3. Hyperscripts † are paired t-test $p < 0.05$ comparing multi-teachers with single-teacher DiSCo. RW denotes the LLM/human rewriting method used as input to the model. FC refers to the models that do not use any rewriting at inference and just take the Full Context as input. ֍ denotes the number of LLM calls used at inference.

	Method	RW	֍	TopiOCQA				QReCC			
				R@100	R@10	MRR	nDCG@3	R@100	R@10	MRR	nDCG@3
Query Rewriting	SPLADE no rewrite	FC	0	0.472	0.258	0.155	0.141	0.840	0.673	0.485	0.459
	SPLADE HumanQR (T_3)	Human	0	-	-	-	-	0.912	0.714	0.448	0.433
	SPLADE T5QR (T_0)	T5	1	0.667	0.501	0.321	0.314	0.840	0.617	0.382	0.366
	SPLADE LlamaQR (T_1)	Llama 3	1	0.761	0.572	0.365	0.352	0.826	0.613	0.377	0.360
	SPLADE MistralQR (T_2)	Mistral 2	1	0.759	0.591	0.366	0.356	0.884	0.668	0.424	0.409
	IterCQR [21]	GPT-3.5	1	0.620	0.426	0.263	0.251	0.841	0.655	0.429	0.402
	LLM4CS [41]	GPT-3.5	5	-	0.433	0.277	0.267	-	0.664	0.448	0.421
	CHIQ FT [43]	T5	1	-	0.510	0.300	0.289	-	0.576	0.369	0.340
	CHIQ-Fusion [43]	Llama 2	6	-	0.616	0.380	$\underline{0.370}$	-	0.707	0.472	0.442
SFT	convANCE [42]	FC	0	0.710	0.430	0.229	0.205	0.872	0.715	0.471	0.456
	convSPLADE [42]	FC	0	0.720	0.521	0.295	0.307	0.878	0.699	0.500	0.466
Distillation	ConvDR [65]	FC	0	0.611	0.435	0.272	0.264	0.778	0.582	0.385	0.357
	QRACDR [46]	FC	0	0.758	0.571	0.377	0.365	0.897	$\underline{0.748}$	**0.516**	**0.516**
	LeCoRe [42]	FC	0	0.735	0.543	0.320	0.314	0.897	0.739	$\underline{0.511}$	0.485
	DiSCo T5	FC	0	0.834	0.617	0.363	0.345	0.917	0.719	0.456	0.442
	DiSCo Mistral	FC	0	$\underline{0.842}$	$\underline{0.634}$	$\underline{0.387}$	$\underline{0.370}$	0.925	0.743	0.489	0.477
	DiSCo Human	FC	0	-	-	-	-	$\underline{0.927}$	0.741	0.483	0.470
	DiSCo multi-teach	FC	0	$\mathbf{0.859}^{\dagger}$	**0.640**	**0.390**	**0.375**	**0.928**	$\mathbf{0.754}^{\dagger}$	$0.498^{\dagger}$	$\underline{0.487}^{\dagger}$

a total limit of 256 tokens for the full conversational context. We also use a 256-token limit for passages. We use a batch size of 10, with 16 negatives per query, and in-batch negatives [35]. Default experiments with SPLADE use FLOPS and L1 regularization, as in the original code, with values $\lambda_q = 1e^{-3}$, $\lambda_d = 5e^{-4}$. We use mixed precision and fp16 to maximize memory use. For sparse retrieval, we use inverted indexes based on the numba library and pyserini [33] together with Pytorch [51]. The fusion method uses the ranx library [5]. All teachers use In-Context learning to generate the rewrites. We provide the one-shot prompt used for rewriting with the LLM Teacher models in Table 2. We experimented in both zero-shot and one-shot, but decided to use one-shot to improve the quality of the teacher models. Given the high efficiency of our fine-tuning, we were able to run the experiments on a single A100 GPU with 40GB of GPU memory.

5 Results

In this section, we present the results of our DiSCo and the proposed distillation for in-domain and out-of-domain retrieval, together with an analysis of the multi-teacher distillation and the efficiency-effectiveness trade-off.

Research questions. We aim to answer the following questions:

RQ1 Would relaxing the distillation training objective improve conversational sparse retrieval in-domain performances?

RQ2 How would the relaxation affect the out-of-domain generalization capacities of the models?

RQ3 How does the distillation of multiple LLM teachers compare to single teacher distillation?

RQ4 How does the relaxed distillation affect the efficiency and sparsity of the student models?

5.1 Performance Comparison

In this first subsection, we aim to answer **RQ1** and **RQ2** by comparing the performance of our proposed DiSCo with state-of-the-art query rewriting, and distillation-based methods. We first compare the performance of DiSCo in the in-domain and out-of-domain retrieval setups in Tables 3 and 4, respectively. The tables report the performance of diverse baselines in terms of various metrics, as well as the number of LLM calls every model requires at inference. This is an important factor while comparing the performance of conversational retrieval models for various reasons, namely, (i) LLM call leads to considerable inference latency, delaying inferencing from hundreds of milliseconds to seconds; and (ii) LLM-based methods take advantage of the vast parameter size and knowledge learned by the training of the LLMs, making their comparison to other smaller models unfair.

In-domain retrieval. Trying to address **RQ1**, we report in Table 3 in-domain performance on TopiOCQA and QReCC. Looking at the results, we do not observe a big gap between the rewriting and distillation-based methods, even though the rewriting-based methods make use of the LLM knowledge in the rewriting phase. Distillation-based methods (i.e., ConvDR, QRACDR, LeCoRe) even

Table 4: Zero-shot Performance on Out-Of-Domain. DiSCo multi-teach uses both T_2 and T_3 (Mistral and Human teachers). DiSCo Fusion is the fusion of the SPLADE MistralQR with DiSCo multi-teach. FC refers to the models that do not use any rewriting at inference and just take the Full Context as input. ֍ denotes the number of LLM calls used at inference.

	Method	RW	֍	CAsT 2020			CAsT 2022			iKAT 2023		
				R@100	MRR	nDCG@3	R@100	MRR	nDCG@3	R@100	MRR	nDCG@3
Query Rewriting	SPLADE HumanQR	Human	0	0.646	0.636	0.475	0.422	0.590	0.423	0.285	0.359	0.262
	SPLADE T5QR	T5	1	0.479	0.477	0.332	0.226	0.355	0.218	0.115	0.200	0.132
	SPLADE LlamaQR	Llama 3	1	0.550	0.515	0.376	0.312	0.453	0.300	0.198	0.281	0.177
	SPLADE MistralQR	Mistral 2	1	0.572	0.553	0.403	0.337	0.487	0.298	0.178	0.291	**0.194**
	LLM4CS [41]	GPT 3.5	5	0.489	0.615	**0.455**	-	-	-	-	-	-
	LLM4CS (ours)	GPT 4	5	0.504	**0.618**	0.444	0.283	0.425	0.272	0.133	0.154	0.099
	CHIQ FT [43]	T5	1	-	0.463	0.316	-	-	-	-	-	-
	CHIQ Fusion [43]	Llama 2	6	-	0.540	0.380	-	-	-	-	-	-
	DiSCo Fusion	Mistral 2	1	**0.611**	0.566	0.425	**0.379**	**0.578**	**0.384**	**0.201**	**0.297**	0.192
SFT	convSPLADE (ours)	FC	0	0.446	0.338	0.234	0.274	0.382	0.227	0.101	0.144	0.085
Distillation	QRACDR [46]	FC	0	0.324	0.442	0.303	-	-	-	-	-	-
	LeCoRe [42]	FC	0	0.467	-	0.290	-	-	-	-	-	-
	DiSCo Mistral	FC	0	0.519	0.457	0.341	0.322	0.463	0.287	0.135	0.193	0.126
	DiSCo Human	FC	0	0.523	0.455	0.339	0.314	0.490	0.308	**0.151**	**0.202**	**0.131**
	DiSCo multi-teach	FC	0	**0.531**	**0.483**	**0.353**	**0.334**	**0.512**	**0.322**	0.147	0.192	0.125

outperform the rewriting-based methods while being more efficient too, as they have zero LLM calls.

Besides, our DiSCo outperforms all rewriting- and distillation-based baselines with a large margin, showing that relaxing the distillation constraint to learn the similarity score, rather than the representations leads to further improvements. In particular, we see that DiSCo Mistral manages to outperform the learned sparse baseline, LeCoRe, by 10 and 2.8 points on TopiOCQA and QReCC resp. in terms of Recall@100.

Out-of-domain retrieval. To further study the effectiveness of our proposed distillation approach, and study its generalizability, we report in Table 4 the results in the out-of-domain retrieval setting on TREC CAsT 20,22 and iKAT 23. Addressing **RQ2**, we see that, unlike the in-domain setting, there is a considerable gap between the rewriting-based and distillation-based methods when it comes to out-of-domain retrieval. On average, rewriting-based methods perform 19% better than distillation-based methods because they mainly rely on the massive parameter size and knowledge of LLMs, leading to high generalizability. However, as mentioned earlier making LLM calls on inference puts the models at a high disadvantage because of the high latency.

Focusing on distillation-based approaches, our DiSCo outperforms other methods by a large margin, showing the remarkable generalizability of our proposed relaxed distillation. In particular, we see that DiSCo Human outperforms QRACDR by 3.8 points in terms of nDCG@3, and LeCoRe by 5.5 points in terms of Recall@100, both on CAsT 2020. Note that those two models are two state-of-the-art models in the same zero-shot settings, trained on QReCC and evaluated on CAsT 2020. It is also noteworthy that even though our distillation-based approach is not able to outperform some of the rewriting-based (LLM-based) methods, our DiSCo-Fusion, which is a fusion method based on LLM rewriting and distillation manages to outperform all rewriting-based methods by a large margin in terms of Recall@100, reaching SotA conversational passage retrieval performance. When comparing closely with the SotA LLM4CS model, we note that while LLM4CS takes advantage of 5 GPT-4 calls, our DiSCo multi-teach model outperforms it in terms of recall on all out-of-domain datasets. Comparing the precision-oriented metrics, we see that DiSCo multi-teach outperforms LLM4CS on all datasets, except TREC CAsT 2020. This dataset was the second edition of TREC CAsT and is considered to be simpler than both 2022 and 2023 versions, showing the performance of our model on complex information needs. Also, considering that we focus on the retrieval task, we consider recall to be preferable, as reranking can be added post-retrieval on a smaller set of documents [11, 62].

Distillation-rewriting fusion. Furthermore, note in Table 4 that DiSCo-Fusion is the fusion of our DiSCo and SPLADE MistralQR, as the fusion of the student with the teacher. Although SPLADE MistralQR exhibits high performance (e.g., 0.57 and 0.33 in terms of Recall@100 resp. on CAsT 2020, 2022), fusing it with the distilled DiSCo leads to further significant improvements (4 points increase on both datasets). This indicates that relaxing the distillation process to the similarity score enables the model to go beyond the knowledge of the teacher in the representation space. This gain is even stronger on the precision of CAsT 2022, where DiSCo-Fusion outperforms SPLADE MistralQR by 8 points in terms of nDCG@3.

Comparison with the teacher models. Now looking at the performance of the students and teachers models in Table 3, we see

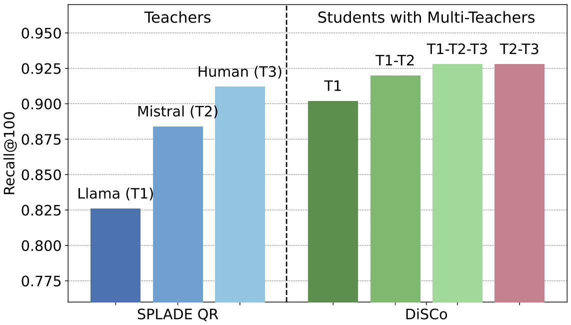

Figure 4: Teacher Selection on QReCC. (Left) SPLADE Teacher Models with different LLM QR. (Right) DiSCo Students when trained with multi-teachers. Best set in red (T_2 and T_3).

that the student models outperform the teachers significantly for in-domain retrieval. In particular, DiSCo Human outperforms SPLADE HumanQR by 3.5 MRR points on QReCC; and similarly with DiSCo Mistral. This demonstrates that our distillation objective allows the student model to learn representations that surpass the original teacher's representations. This is possible thanks to the relaxation of the distillation, as the student can adapt and learn optimal representations based on gradient descent optimization. This was not possible with the previous distillation objective on the representation, as student models were trained to copy the representations of the teacher model. Out-of-domain however, we see in Table 4 that the student models have more difficulties outperforming the performance of the teacher models. This result on out-of-domain retrieval is, however, not fair considering that teacher model generalization relies on billions of parameters, while students on a few hundred. Recall also that DiSCo-Fusion as the fusion of the teacher and the student was further improving performance out-of-domain.

Reliance on weak teacher model. Finally, we explore the reliance on the teacher models, by training DiSCo with a weaker teacher model, such as T5. This is interesting in scenarios where we do not have access to strong teachers and also shows the robustness of the method when trained on more noisy labels. From Table 3, we see that DiSCo T5 performs almost on par with DiSCo Mistral, while SPLADE MistralQR outperforms SPLADE T5QR by an important margin, on both TopiOCQA and QReCC. Our method is thus robust even when trained on lower-quality rewrites from the teacher.

5.2 Multiple Teachers Distillation

While the previous section focused on the distillation of single teachers, here we answer **RQ3** on the use of multiple teachers. Thus, we compare our DiSCo model trained with multiple teachers.

In-domain and out-of-domain retrieval. Trying to address **RQ3**, both Tables 3 and 4 include our DiSCo multi-teach model trained with multiple teachers. Considering the performances of DiSCo multi-teach for in-domain retrieval, we observe significant gains compared to DiSCo Mistral on QReCC. In particular, we observe a 1-point increase in terms of R@10, MRR, and nDCG@3 on QReCC. This result remains consistent for out-of-domain retrieval, where within Table 4 DiSCo multi-teach outperforms its single version model, by at least 2 points on all metrics, on CAsT 2020 and 2022. On iKAT 2023, however, we observe mixed results. Two interpretations are possible for these general gains: first, the multiple teachers and their subsequent rewrites produce word synonyms, as a form of query expansion, improving the performance, and second, as one LLM may fail to resolve the context modeling task, another can balance it, offering better robustness.

Table 5: Inference Efficiency on TREC CAsT 2020. Rewrite and query encoding are measured on GPU, while the inverted index search on CPU with the Numba library. DiSCo Fusion is the fusion of SPLADE MistralQR with DiSCo Mistral.

Efficiency (ms)	Rewrite	Retrieval	Total	R@100
SPLADE T5QR	190	358	548	0.479
SPLADE LlamaQR	4245	367	4612	0.550
SPLADE MistralQR	2094	356	2450	0.572
convSPLADE	0	356	356	0.446
DiSCo Mistral	0	346	346	0.519
DiSCo multi-teach	0	357	357	0.531
DiSCo Fusion	2094	356	**2450**	**0.611**

Progressive Improvement. Further addressing **RQ3**, Figure 4 gives insights on the multiple teachers distillation. While the left part of the Figure gives the performances of the teachers on QReCC, the right part shows the gains when incrementally adding teachers to the distillation. We can also notice the important gains from the distillation of LlamaQR. In particular, DiSCo Llama improves by 8 points compared to SPLADE LlamaQR in terms of Recall@100, while in comparison DiSCo Mistral by only 4 points compared to his teacher. This shows that the relaxation is robust and can learn from even lower-quality rewrites.

5.3 Effectiveness-Efficiency Trade-off

In this subsection, we aim to answer **RQ4** by analyzing the effectiveness-efficiency trade-off of the proposed DiSCo models. First, we measure the efficiency of our method compared to rewriting-based methods. Then, we show the possibility of controlling the sparsity of the model through the FLOP regularization. Finally, we dive into sparsity at different depths of conversations.

Efficiency. As to explicit the inference efficiency gain of our method compared to rewrite-based approaches, we plot in Table 5 the efficiency in milliseconds of the models on TREC CAsT 2020. We observe from the table that while SPLADE MistralQR achieves very high performance here out-of-domain, the efficiency is very low compared to DiSCo Mistral. This is because of the rewriting step, which involves 1 LLM call for MistralQR. Also note that the rewrite needs to be executed on GPU, while most of the retrieval time is a search through the inverted index on CPU. DiSCo Mistral and multi-teach thus appear to have a better trade-off when it comes to efficiency. Finally, DiSCo Fusion has the best performance overall, without an important computational overload compared

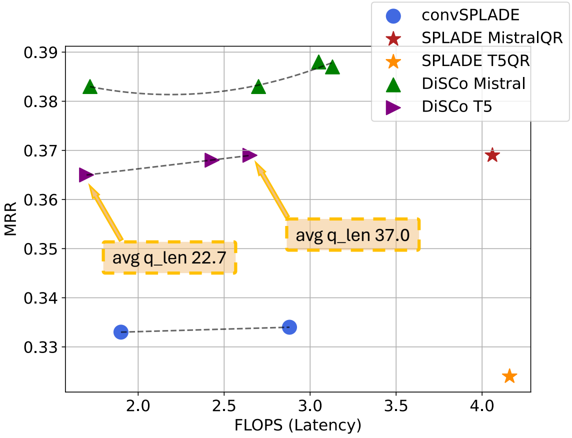

Figure 5: Effectiveness-efficiency trade-off on TopiOCQA. Sparser representations have a lower latency but also a lower MRR. `avg q_len` is the average number of activated tokens in the conversation representations, as indicator of efficiency.

to SPLADE MistralQR [5]. Note that retrieval is not optimized, and libraries such as PISA could reduce retrieval of SPLADE models below 100 ms on CPU, as shown in recent papers [27, 30, 32]. This would further increase the efficiency gap between rewrite-based and distillation-based approaches. We also focus here on inference efficiency, as training efficiency does not directly affect the user, and would involve training LLMs. Those results provide a first answer to **RQ4** on the efficiency of DiSCo compared to existing methods.

Controlling Sparsity. We then focus here on the sparsity of the sparse representations, as a measure of efficiency within the inverted index. Figure 5 plots the performance and FLOPs of several DiSCo models when trained with different levels of sparsity. This is possible thanks to the regularization loss from the SPLADE architecture, controlled by the hyper-parameters (λ_q, λ_d) [17]. From the Figure, we see for example that the same DiSCo T5 model can be trained with different degrees of regularization, leading to different FLOPs and MRR. Furthermore, Figure 5 shows the FLOPs of the associated teacher models: SPLADE MistralQR and T5QR. We notice that our student DiSCo models are more sparse compared to their teachers, as having only a constraint on similarities during training allows more freedom on the representations and their sparsity. This answers part of **RQ4**, showing that DiSCo can be trained with different sparsity levels.

Sparsity at different Conversation Depths. To further address **RQ4**, we plot the sparsity and effectiveness of several baselines in Figure 6. On the left part of the graph, we see the performance according to the depth of the conversation – as the number of previous interactions between the user and the system – and on the right the number of activated tokens in the representations. In particular, we see the difficulty for long conversations across models. Comparing our DiSCo with the convSPLADE baseline, we see that the main improvement comes from longer conversations.

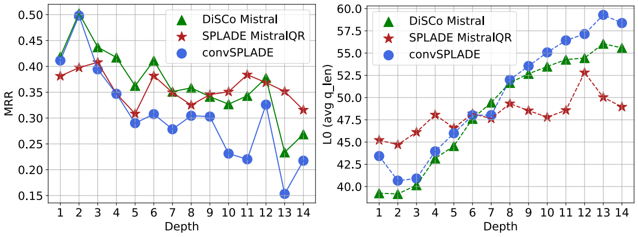

Figure 6: (Left) Performance with respect to depth on TopiOCQA. (Right) Sparsity of query representations with respect to depth of conversations.

We also see that DiSCo outperforms its teacher model, SPLADE MistralQR, on the first few turns of the conversation. This is also a difficult task here as the TopiOCQA dataset has topic switches every few turns, making the context modeling task more complex.

Examining sparsity, we observe that the representations across all models remain sparse, with at most 60 non-zero dimensions even at deeper conversation levels (> 10). At these depths, conversations become significantly longer, reaching up to 350 tokens. This suggests that the models effectively perform noise reduction as part of the context modeling task, activating only a subset of the tokens from the context. Now, comparing the models, as SPLADE MistralQR relies on Mistral rewrite, we see that the representation sparsity is consistent even for long conversations. For DiSCo and convSPLADE, sparsity decreases with the turns of the conversation.

6 Conclusion & Future Work

In this work, we propose DiSCo, a novel distillation strategy in CS to distill query rewriting from LLMs. DiSCo trains CS models with a relaxed distillation strategy that unifies context modeling and retrieval tasks. Our experiments on LSR and DiSCo models demonstrate that this training objective achieves important performance gains in both in-domain and out-of-domain retrieval. By distilling LLM rewrites, our method effectively learns from a single or several LLM teachers, outperforming the teachers, and reaching SoTA on several CS datasets. The proposed training strategy of DiSCo also shows robustness to the quality of the teacher model when trained on weaker teacher models. We further examine the inference efficiency and sparsity of our approach after distillation. Our findings emphasize the importance of aligning the rewriting and retrieval tasks in CS, with training objectives that unify both tasks. As a future work, we plan to study different teachers, as any model producing a similarity score could be distilled, e.g., cross-encoder [20], paired with stronger LLMs for rewrite.

Acknowledgments

This research was partly supported by the Swiss National Science Foundation (SNSF), under the project PACINO (Personality And Conversational INformatiOn Access), grant number 215742. This funding is a collaboration between the University of Amsterdam and the Università della Svizzera italiana.

[5] DiSCo Fusion being an ensembling of two approaches, we report the maximum execution time of the two (as they could be executed in parallel).

References

[1] Zahra Abbasiantaeb, Simon Lupart, and Mohammad Aliannejadi. 2024. Generating Multi-Aspect Queries for Conversational Search. *arXiv preprint arXiv:2403.19302* (2024).

[2] Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. TopiOCQA: Open-domain Conversational Question Answering with Topic Switching. *Transactions of the Association for Computational Linguistics* 10 (2022), 468–483. doi:10.1162/tacl_a_00471

[3] Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffrey Dalton, and Leif Azzopardi. 2024. TREC iKAT 2023: A Test Collection for Evaluating Conversational and Interactive Knowledge Assistants. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 819–829. doi:10.1145/3626772.3657860

[4] Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-Domain Question Answering Goes Conversational via Question Rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 520–534. doi:10.18653/v1/2021.naacl-main.44

[5] Elias Bassani. 2022. ranx: A Blazing-Fast Python Library for Ranking Evaluation and Comparison. In *ECIR (2) (Lecture Notes in Computer Science, Vol. 13186)*. Springer, 259–264. doi:10.1007/978-3-030-99739-7_30

[6] B. Barla Cambazoglu, Leila Tavakoli, Falk Scholer, Mark Sanderson, and Bruce Croft. 2021. An Intent Taxonomy for Questions Asked in Web Search. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (Canberra ACT, Australia) *(CHIIR '21)*. Association for Computing Machinery, New York, NY, USA, 85–94. doi:10.1145/3406522.3446027

[7] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC : Question Answering in Context. *CoRR* abs/1808.07036 (2018). arXiv:1808.07036 http://arxiv.org/abs/1808.07036

[8] Jeffrey Dalton, Sophie Fischer, Paul Owoicho, Filip Radlinski, Federico Rossetto, Johanne R. Trippas, and Hamed Zamani. 2022. Conversational Information Seeking: Theory and Application. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 3455–3458. doi:10.1145/3477495.3532678

[9] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. CAsT 2020: The Conversational Assistance Track Overview. In *Text Retrieval Conference*. https://api.semanticscholar.org/CorpusID:214735659

[10] Jeffrey Dalton, Chenyan Xiong, Vaibhav Kumar, and Jamie Callan. 2020. CAsT-19: A Dataset for Conversational Information Seeking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 1985–1988. doi:10.1145/3397271.3401206

[11] Hervé Déjean, Stéphane Clinchant, and Thibault Formal. 2024. A Thorough Comparison of Cross-Encoders and LLMs for Reranking SPLADE. *arXiv preprint arXiv:2403.10407* (2024).

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[13] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, and et al. 2024. The Llama 3 Herd of Models. *CoRR* abs/2407.21783 (2024). doi:10.48550/ARXIV.2407.21783 arXiv:2407.21783

[14] Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can You Unpack That? Learning to Rewrite Questions-in-Context. 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)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 5918–5924. doi:10.18653/v1/D19-1605

[15] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086* (2021).

[16] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. From Distillation to Hard Negative Sampling: Making Sparse Neural IR Models More Effective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2353–2359. doi:10.1145/3477495.3531857

[17] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 2288–2292. doi:10.1145/3404835.3463098

[18] Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2022. Neural Approaches to Conversational Information Retrieval. *CoRR* abs/2201.05176 (2022). arXiv:2201.05176 https://arxiv.org/abs/2201.05176

[19] Nam Hai Le, Thomas Gerald, Thibault Formal, Jian-Yun Nie, Benjamin Piwowarski, and Laure Soulier. 2023. CoSPLADE: Contextualizing SPLADE for Conversational Information Retrieval. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I* (Dublin, Ireland). Springer-Verlag, Berlin, Heidelberg, 537–552. doi:10.1007/978-3-031-28244-7_34

[20] Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020. Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv preprint arXiv:2010.02666* (2020).

[21] Yunah Jang, Kang-il Lee, Hyunkyung Bae, Hwanhee Lee, and Kyomin Jung. 2024. IterCQR: Iterative Conversational Query Reformulation with Retrieval Guidance. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 8121–8138. doi:10.18653/v1/2024.naacl-long.449

[22] Kalervo Järvelin and Jaana Kekäläinen. 2000. IR Evaluation Methods for Retrieving Highly Relevant Documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Athens, Greece) *(SIGIR '00)*. ACM, New York, NY, USA, 41–48. doi:10.1145/345508.345545

[23] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

[24] Gregory R. Koch. 2015. Siamese Neural Networks for One-Shot Image Recognition. https://api.semanticscholar.org/CorpusID:13874643

[25] Solomon Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *Annals of Mathematical Statistics* 22 (1951), 79–86. https://api.semanticscholar.org/CorpusID:120349231

[26] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: a Benchmark for Question Answering Research. *Transactions of the Association of Computational Linguistics* (2019).

[27] Carlos Lassance and Stéphane Clinchant. 2022. An Efficiency Study for SPLADE Models. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2220–2226. doi:10.1145/3477495.3531833

[28] Carlos Lassance and Stéphane Clinchant. 2022. Naver Labs Europe (SPLADE) @ TREC NeuCLIR 2022. In *TREC*. https://trec.nist.gov/pubs/trec31/papers/NLE.N.pdf

[29] Carlos Lassance and Stéphane Clinchant. 2023. Naver Labs Europe (SPLADE)@ TREC Deep Learning 2022. *arXiv preprint arXiv:2302.12574* (2023).

[30] Carlos Lassance, Hervé Dejean, Stéphane Clinchant, and Nicola Tonellotto. 2024. Two-Step SPLADE: Simple, Efficient and Effective Approximation of SPLADE. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part II* (Glasgow, United Kingdom). Springer-Verlag, Berlin, Heidelberg, 349–363. doi:10.1007/978-3-031-56060-6_23

[31] Carlos Lassance, Hervé Déjean, Thibault Formal, and Stéphane Clinchant. 2024. SPLADE-v3: New baselines for SPLADE. *arXiv preprint arXiv:2403.06789* (2024).

[32] Carlos Lassance, Simon Lupart, Hervé Déjean, Stéphane Clinchant, and Nicola Tonellotto. 2023. A Static Pruning Study on Sparse Neural Retrievers. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 1771–1775. doi:10.1145/3539618.3591941

[33] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*. 2356–2362.

[34] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020. Pretrained Transformers for Text Ranking: BERT and Beyond. *arXiv:2010.06467 [cs]* (Oct. 2020). http://arxiv.org/abs/2010.06467 ZSCC: NoCitationData[s0] arXiv: 2010.06467.

[35] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. In-Batch Negatives for Knowledge Distillation with Tightly-Coupled Teachers for Dense Retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*. Association for Computational Linguistics, Online, 163–173. doi:10.18653/v1/2021.repl4nlp-1.17

[36] Zhiyuan Liu, Yankai Lin, and Maosong Sun. 2021. Representation Learning for Natural Language Processing. *CoRR* abs/2102.03732 (2021). arXiv:2102.03732 https://arxiv.org/abs/2102.03732

[37] Simon Lupart, Zahra Abbasiantaeb, and Mohammad Aliannejadi. 2024. IRLab@ iKAT24: Learned Sparse Retrieval with Multi-aspect LLM Query Generation for Conversational Search. *arXiv preprint arXiv:2411.14739* (2024).

[38] Simon Lupart and Stéphane Clinchant. 2023. A Study on FGSM Adversarial Training for Neural Retrieval. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part II* (Dublin, Ireland). Springer-Verlag, Berlin, Heidelberg, 484–492. doi:10.1007/978-3-031-28238-6_39

[39] Simon Lupart, Thibault Formal, and Stéphane Clinchant. 2023. Ms-shift: An analysis of ms marco distribution shifts on neural retrieval. In *European Conference on Information Retrieval*. Springer, 636–652.

[40] Kelong Mao, Zhicheng Dou, Bang Liu, Hongjin Qian, Fengran Mo, Xiangli Wu, Xiaohua Cheng, and Zhao Cao. 2023. Search-Oriented Conversational Query Editing. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 4160–4172. doi:10.18653/v1/2023.findings-acl.256

[41] Kelong Mao, Zhicheng Dou, Fengran Mo, Jiewen Hou, Haonan Chen, and Hongjin Qian. 2023. Large Language Models Know Your Contextual Search Intent: A Prompting Framework for Conversational Search. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 1211–1225. doi:10.18653/v1/2023.findings-emnlp.86

[42] Kelong Mao, Hongjin Qian, Fengran Mo, Zhicheng Dou, Bang Liu, Xiaohua Cheng, and Zhao Cao. 2023. Learning Denoised and Interpretable Session Representation for Conversational Search. In *Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) *(WWW '23)*. Association for Computing Machinery, New York, NY, USA, 3193–3202. doi:10.1145/3543507.3583265

[43] Fengran Mo, Abbas Ghaddar, Kelong Mao, Mehdi Rezagholizadeh, Boxing Chen, Qun Liu, and Jian-Yun Nie. 2024. CHIQ: Contextual History Enhancement for Improving Query Rewriting in Conversational Search. *arXiv preprint arXiv:2406.05013* (2024).

[44] Fengran Mo, Kelong Mao, Ziliang Zhao, Hongjin Qian, Haonan Chen, Yiruo Cheng, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, and Jian-Yun Nie. 2024. A survey of conversational search. *arXiv preprint arXiv:2410.15576* (2024).

[45] Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. 2023. ConvGQR: Generative Query Reformulation for Conversational Search. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 4998–5012. doi:10.18653/v1/2023.acl-long.274

[46] Fengran Mo, Chen Qu, Kelong Mao, Yihong Wu, Zhan Su, Kaiyu Huang, and Jian-Yun Nie. 2024. Aligning Query Representation with Rewritten Query and Relevance Judgments in Conversational Search. *arXiv preprint arXiv:2407.20189* (2024). arXiv:2407.20189 [cs.IR] https://arxiv.org/abs/2407.20189

[47] 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. *CoRR* abs/1611.09268 (2016). arXiv:1611.09268 http://arxiv.org/abs/1611.09268

[48] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[49] Paul Owoicho, Jeffrey Dalton, Mohammad Aliannejadi, Leif Azzopardi, Johanne R. Trippas, and Svitlana Vakulenko. 2022. TREC CAsT 2022: Going Beyond User Ask and System Retrieve with Initiative and Response Generation. In *Text Retrieval Conference*. https://api.semanticscholar.org/CorpusID:261288646

[50] Biswajit Paria, Chih-Kuan Yeh, Ian En-Hsu Yen, Ning Xu, Pradeep Ravikumar, and Barnabás Póczos. 2020. Minimizing FLOPs to Learn Efficient Sparse Representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=SygpC6Ntvr

[51] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library.. In *NeurIPS*.

[52] Gustavo Penha and Claudia Hauff. 2020. Challenges in the Evaluation of Conversational Search Systems.. In *Converse@ KDD*.

[53] Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W. Bruce Croft, and Mohit Iyyer. 2020. Open-Retrieval Conversational Question Answering. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 539–548. doi:10.1145/3397271.3401110

[54] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (Oslo, Norway) *(CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 117–126. doi:10.1145/3020165.3020183

[55] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *CoRR* abs/1910.10683 (2019). arXiv:1910.10683 http://arxiv.org/abs/1910.10683

[56] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. 2015. Deep Metric Learning via Lifted Structured Feature Embedding. *CoRR* abs/1511.06452 (2015). arXiv:1511.06452 http://arxiv.org/abs/1511.06452

[57] Tao Tao and ChengXiang Zhai. 2007. An exploration of proximity measures in information retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Amsterdam, The Netherlands) *(SIGIR '07)*. Association for Computing Machinery, New York, NY, USA, 295–302. doi:10.1145/1277741.1277794

[58] Tongzhou Wang and Phillip Isola. 2020. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. *arXiv preprint arXiv:2005.10242* (2020).

[59] Kilian Q. Weinberger and Lawrence K. Saul. 2009. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *J. Mach. Learn. Res.* 10 (jun 2009), 207–244.

[60] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Qun Liu and David Schlangen (Eds.). Association for Computational Linguistics, Online, 38–45. doi:10.18653/v1/2020.emnlp-demos.6

[61] Zeqiu Wu, Yi Luan, Hannah Rashkin, D. Reitter, and Gaurav Singh Tomar. 2021. CONQRR: Conversational Query Rewriting for Retrieval with Reinforcement Learning. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:245218563

[62] Yan Xiao, Yixing Fan, Ruqing Zhang, and Jiafeng Guo. 2023. Beyond Precision: A Study on Recall of Initial Retrieval with Neural Representations. In *Information Retrieval: 28th China Conference, CCIR 2022, Chongqing, China, September 16–18, 2022, Revised Selected Papers* (Chongqing, China). Springer-Verlag, Berlin, Heidelberg, 76–89. doi:10.1007/978-3-031-24755-2_7

[63] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. *CoRR* abs/2007.00808 (2020). arXiv:2007.00808 https://arxiv.org/abs/2007.00808

[64] Fanghua Ye, Meng Fang, Shenghui Li, and Emine Yilmaz. 2023. Enhancing Conversational Search: Large Language Model-Aided Informative Query Rewriting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 5985–6006. doi:10.18653/v1/2023.findings-emnlp.398

[65] Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-Shot Conversational Dense Retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 829–838. doi:10.1145/3404835.3462856

[66] Hamed Zamani, Mostafa Dehghani, W. Bruce Croft, Erik Learned-Miller, and Jaap Kamps. 2018. From Neural Re-Ranking to Neural Ranking: Learning a Sparse Representation for Inverted Indexing. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (Torino, Italy) *(CIKM '18)*. Association for Computing Machinery, New York, NY, USA, 497–506. doi:10.1145/3269206.3271800