SSP: Self-Supervised Post-training for Conversational Search

Quan Tu[1*], **Shen Gao**[2*], **Xiaolong Wu**[4], **Zhao Cao**[4], **Ji-Rong Wen**[1,3], **Rui Yan**[1,3†]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]School of Computer Science and Technology, Shandong University
[3]Engineering Research Center of Next-Generation Intelligent
Search and Recommendation, Ministry of Education
[4]Huawei Poisson Lab
[1]{quantu,jrwen,ruiyan}@ruc.edu.cn, [2]shengao@pku.edu.cn
[4]{wuxiaolong19, caozhao1}@huawei.com

Abstract

Conversational search has been regarded as the next-generation search paradigm. Constrained by data scarcity, most existing methods distill the well-trained ad-hoc retriever to the conversational retriever. However, these methods, which usually initialize parameters by query reformulation to discover contextualized dependency, have trouble in understanding the dialogue structure information and struggle with contextual semantic vanishing. In this paper, we propose **S**elf-**S**upervised **P**ost-training (SSP) which is a new post-training paradigm with three self-supervised tasks to efficiently initialize the conversational search model to enhance the dialogue structure and contextual semantic understanding. Furthermore, the SSP can be plugged into most of the existing conversational models to boost their performance. To verify the effectiveness of our proposed method, we apply the conversational encoder post-trained by SSP on the conversational search task using two benchmark datasets: CAsT-19 and CAsT-20. Extensive experiments that our SSP can boost the performance of several existing conversational search methods. Our source code is available at `https://github.com/morecry/SSP`.

1 Introduction

The past years have witnessed the fast progress of the ad-hoc search(Dai and Callan, 2020; Dai et al., 2018; Fujiwara et al., 2013; Gao et al., 2019). However, when it confronts more complicated information needs, the traditional ad-hoc search seems to be less competent. Recently, researchers proposes conversational search which is the combination of the search engine and the conversational assistant (Radlinski and Craswell, 2017; Zhang et al., 2018; Kiesel et al., 2021; Trippas et al., 2020; Tu et al., 2022). Different from the keyword-based query in the ad-hoc search, multi-turn natural language utterance is the main interactive form in the conversational search. This yields the challenge of developing the conversational search system that existing ad-hoc retrievers and datasets cannot be directly used to derive the conversational query understanding module.

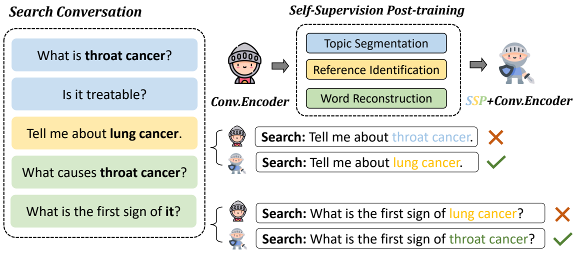
Search Conversation
What is **throat cancer**?
Is it treatable?
Tell me about **lung cancer**.
What causes **throat cancer**?
What is the first sign of **it**?
Conv.Encoder
Self-Supervision Post-training
Topic Segmentation
Reference Identification
Word Reconstruction
SSP+Conv.Encoder
Search: Tell me about throat cancer. ✗
Search: Tell me about lung cancer. ✓
Search: What is the first sign of lung cancer? ✗
Search: What is the first sign of throat cancer? ✓

Figure 1: Example of modeling the conversational structure in conversational search. The model should capture the structure including the topic has been shifted at the 3rd utterance and the last utterance has coreference with the previous utterance. This information can help the model understand the search intent of users accurately.

In the beginning, researchers reformulate a conversational query to a de-contextual query, which is used to perform ad-hoc retrieval (Lin et al., 2020b; Mele et al., 2021; Lin et al., 2021b). Recently the conversational dense retrieval model (Lin et al., 2021a; Mao et al., 2022) is presented to directly encode the whole multi-turn conversational context as a vector representation and conduct matching with the candidate document representations. Since the real-world conversational search corpus is hard to collect, a warm-up step is additionally employed to initialize the conversational representation ability (Yu et al., 2021; Dai et al., 2022). These conversational dense retrieval methods have achieved significantly better performance than the query reformulation methods and have been widely adopted in research of conversational search (Yu et al., 2021; Dai et al., 2022). However, these warm-up methods just use the same training objective on a large dataset from other domains to initialize the

[*] Equal Contribution.
[†] Corresponding author: Rui Yan (ruiyan@ruc.edu.cn).

Findings of the Association for Computational Linguistics: ACL 2023, pages 13237–13249
July 9-14, 2023 ©2023 Association for Computational Linguistics

parameters of the conversational encoder, which can hardly capture the structure information of the conversation which is essential for understanding the user's search intent accurately.

In this paper, we propose **S**elf-**S**upervised **P**ost-training (SSP) for the conversational search task as shown in Figure 1. In SSP, we replace the commonly used warm-up step with a new post-training paradigm which contains three novel self-supervised tasks to learn how to capture the structure information and keep contextual semantics. To be more specific, the first self-supervised task is *topic segmentation*, which learns to decompose the dialogue structure into several segments based on the topic. To tackle the coreference problem which is a ubiquitous problem of multi-turn conversation modeling, we propose the *coreference identification* task which helps the model identify the most possible referred terms in the context and simplifies the intricate dialogue structure. Since understanding and remembering the semantic information in the conversational context is vital for conversational context modeling, we propose the *word reconstruction* task which prevents contextual semantic vanishing. To demonstrate the effectiveness of SSP, we first equip several existing conversational search methods with SSP and conduct experiments on two benchmark datasets: CAsT-19 (Dalton et al., 2020) and CAsT-20 (Dalton et al., 2021). Experimental results demonstrate that the SSP outperforms all the strong baselines on 2 datasets.

To sum up, our contributions can be summarized as follows:

- We propose a general and extensible post-training framework to better initialize the conversational context encoder in the existing conversational search models.
- We propose three specific self-supervised tasks which help the model to capture the conversational structure information and prevent the contextual semantics from vanishing.
- Experiments show that our SSP can boost the performance of strong conversational search methods on two benchmark datasets and achieves state-of-the-art performance.

2 Related Work

Conversational search has become a hop research topic in recent years. TREC Conversational Assistant Track (CAsT) competition (Dietz et al., 2017), which holds the benchmark largely promotes the progress of conversational search. In the beginning, researchers simply view conversational search as the query reformulation problem. They suppose that if a context-dependent query could be rewritten to a de-contextualized query based on historical queries, then it directly uses the well-trained ad-hoc retriever to obtain retrieval results. Transformer++ (Vakulenko et al., 2021) fine-tunes the GPT-2 on query reformulation dataset CANARD (Elgohary et al., 2019) to rewrite query. QueryRewriter (Yu et al., 2020) exploits large amounts of ad-hoc search sessions to build a weak-supervision query reformulation data generator, then these automatically generated data is used to fine-tune the language model. However, these methods underestimate the value of context, which contains various latent search intentions and topic information.

After that, the conversational dense retriever is proposed. It straightly encodes full conversation whose last query denotes the user's real search intention to dense representation. ConvDR (Yu et al., 2021) forces the contextual representation to mimic the reformulation query representation based on the teacher-student framework, which slightly deals with the conversational search data scarcity problem. Further, COTED (Mao et al., 2022) points out that not all queries in context are useful and devises a curriculum denoising method to inhibit the influence of unnecessary contextual queries. These dense methods additionally perform the warm-up on the other domain dataset to initialize the parameters based on their own objective. However, their warm-up ignore the conversation structure information, which is crucial for capturing the relationship between utterances and understanding the search intention of the user. In this respect, we devise a novel **S**elf-**S**upervised **P**ost-training (SSP) to replace the warm-up as Figure 2.

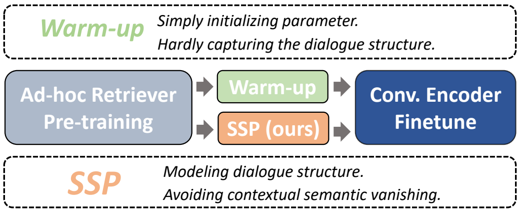
Warm-up
Simply initializing parameter.
Hardly capturing the dialogue structure.
Ad-hoc Retriever Pre-training
Warm-up
SSP (ours)
Conv. Encoder Finetune
SSP
Modeling dialogue structure.
Avoiding contextual semantic vanishing.

Figure 2: The comparison between the training procedure of conversational search with warm-up and the SSP paradigm.

3 Problem formulation

We assume that there is a multi-turn search conversation $Q = \{q_1, q_2, \dots, q_n\}$, where $q_i = \{x_{i,1}, x_{i,2}, \dots, x_{i,l_i}\}$ represents the i-th question in the conversation and $x_{i,j}$ is the j-th token in q_i. The last query q_n is the user's real search intention. We insert special tokens [CLS] and [SEP] in Q yielding $\{\texttt{CLS}, q_1, [\texttt{SEP}], q_2, [\texttt{SEP}], \dots, [\texttt{SEP}], q_n\}$ as the model input, where [CLS] is the start token and [SEP] is the separation token to split each query. After the concatenation of all queries is sent into the conversational encoder (a transformer-based architecture model), we obtain the last layer's output hidden state E. $E_{[\texttt{CLS}]}$ and $E_{[\texttt{SEP}]}$ are the corresponding representations of [CLS] and [SEP] and will be used in self-supervised tasks. Our goal is to learn a better contextual representation $E_{[\texttt{CLS}]}$ in order to accurately retrieve documents in corpus for the last query q_n.

4 Self-Supervised Post-training

4.1 Overview

In this section, we propose our **S**elf-**S**upervised **P**ost-training, abbreviated as SSP. An overview of SSP is shown in Figure 3, which consists of three self-supervised tasks:

- *Topic Segmentation Task* aims to find the topic-shifting point in the utterances. It helps the model to capture the topic structure in the conversational context.
- *Coreference Identification Task* aims to identify the correlation structure between two referred utterances, which helps the conversational encoder to understand the coreference relationship and produce better query representation.
- *Word Reconstruction Task* aims to reconstruct the bag-of-word (BOW) vector of the conversational context using the conversational vector representation. It helps the model avoid the contextual semantic vanishing during conversation encoding.

After jointly training the conversational encoder using these three self-supervised tasks, we fine-tune the encoder to the conversational search downstream task using the existing conversational search methods.

4.2 Topic Segmentation Task

When the user interacts with the conversational search system, the focused topic may vary from time to time. Taking the example in Figure 1, the search intention of the user changes according to the retrieval results of previous turns. This causes the topic of the conversation to shift. Since the conversation topic may shift in every utterance, to fully understand a user query, the conversational system should know what is the current topic of this query and view the utterances of the current topic as a more salient context. If the conversational encoder cannot identify the topic boundary of the current topic, it may focus on unrelated utterances and incorporate noise information into the query representation.

Thus we propose the topic segmentation task to identify the topic boundary of the conversation, which can facilitate the model to focus on more related context when encoding the query. We first randomly sample a noise conversational session with several utterances from the training corpus and then concatenate this sampled noise session at the beginning of the raw conversational context. Given the raw search conversation $Q = \{q_1, q_2, \dots, q_n\}$ and the noisy conversation $Q' = \{q'_1, q'_2, \dots, q'_m\}$, we truncate the first k queries of Q' where k is sampled based on reciprocal probability distribution p, which avoids the distortion of the raw context from the abundant long noisy sessions,

$$p_k = \frac{1}{k} / \sum_{i=1}^{m} \frac{1}{i}, k = 1, 2, \dots, m.$$

After concatenating the sampled noise session before the raw context and separating each query by [SEP], we obtain the perturbed conversation $\check{Q} = \{[\texttt{CLS}], q'_1, [\texttt{SEP}], \dots, q'_k, [\texttt{SEP}], q_1, [\texttt{SEP}], \dots, q_n\}$ and the ground truth topic label $y^t = \{1, \dots, 1, 0, \dots, 0\}$, where the queries from the external conversation are labelled as 1 and the ones from the raw conversation are labelled as 0.

Next, we use the perturbed conversation $\check{Q}$ as input to the conversational encoder, and obtain the vector representation $\check{E} = \{E_{[\texttt{CLS}]}, E'_1, E_{[\texttt{SEP}]}, \dots, E'_k, E_{[\texttt{SEP}]}, E_1, E_{[\texttt{SEP}]}, \dots, E_n\}$ of the perturbed conversation $\check{Q}$. Finally, $E_{[\texttt{SEP}]}$ is sent to the topic predictor (a linear layer) to decide whether an utterance is from the sampled noise conversation Q' or not. The binary cross entropy is used to compute topic segmentation loss $\mathcal{L}_{TS}$:

$$\begin{aligned} p(y_i^t = 1|\check{Q}) &= \text{Sigmoid}(W_t E_{[\texttt{SEP}]} + b_t), \\ \mathcal{L}_{TS} &= -y_i^t \log(p(y_i^t = 1|\check{Q})) \\ &\quad - (1 - y_i^t)(1 - \log(p(y_i^t = 1|\check{Q}))), \end{aligned}$$

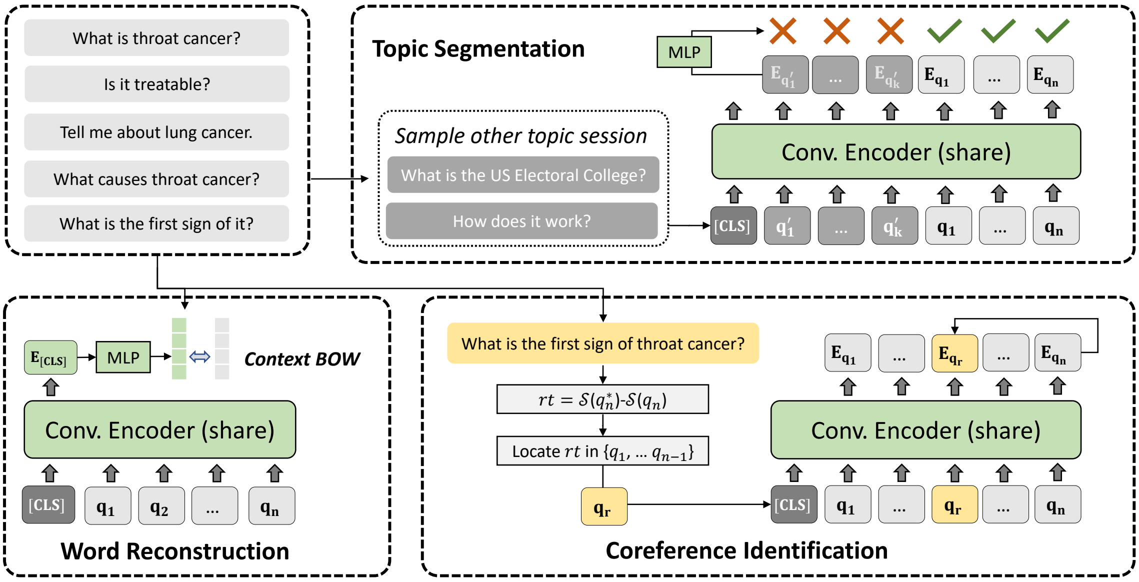

Figure 3: Overview of SSP. It consists of three self-supervised tasks to conduct post-training of conversational encoder: (1) *Topic Segmentation* predicts which utterances are the randomly sampled perturbation utterances from other conversation sessions; (2) *Coreference Identification* predicts which utterance in the conversational context is related to the last utterance; (3) *Word Reconstruction* uses the conversational context vector representation to reconstruct the Bag-of-Word vector of conversational context.

where $W_t \in \mathbf{R}^{h \times 1}, b_t \in \mathbf{R}$, h is the hidden size of model.

4.3 Coreference Identification Task

In conversational search, a common problem is the coreference, which is that the pronoun in a query usually refers to a term in its previous queries. Most of the existing methods did not explicitly train the model to tackle this problem. Here, we devise an auxiliary self-supervised task that trains the model to predict the referred utterance of the last utterance by the coreference relationship. To obtain which utterance in the conversational context has the coreference relationship with the last utterance, we use the query reformulation corpus to find. We compare the last query in Q with the reformulated query q_n^* by set operations to find the reformulation terms r have been omitted in Q:

$$r = \mathcal{S}(\text{tokenize}(q_n^*)) - \mathcal{S}(\text{tokenize}(q_n)),$$

where $\mathcal{S}$ is a set operation that converts a sentence into a non-repeating word set. We can obtain the reformulation terms r by calculating the difference set between two sets. Then r will be used to locate the referred query from back to front until the first query containing the r is found. We mark the position of the referred query to the label $y^c = \{0, 0, \ldots, 1, \ldots, 0\}$, whose i-th value is 1 only if the i-th query is the referred query. Similar to the topic segmentation task (introduced in § 4.2), we send $E_{\text{[SEP]}}$ into a coreference predictor to predict the referred query and use the binary cross-entropy as the loss function of this task:

$$\begin{aligned} p(y_i^c = 1|Q) &= \text{Sigmoid}(W_c E_{\text{[SEP]}} + b_c), \\ \mathcal{L}_{CI} &= -y_i^c \log(p(y_i^c = 1|Q)) \\ &\quad - (1 - y_i^c)(1 - \log(p(y_i^c = 1|Q))), \end{aligned}$$

where $W_r \in \mathbf{R}^{h \times 1}, b_r \in \mathbf{R}$ are all trainable parameters. With the coreference identification task, the conversational encoder will pay more attention to the most possible referred query in context when it understands the last query.

4.4 Word Reconstruction Task

The duality of a one-stage conversational retriever will encode a query to a dense vector. In the previous sections, we use the self-supervised tasks to focus on the utterance of the current topic and the highly related utterance with coreference. However, other utterances may also provide useful information to understand the current search intent. Thus, the conversational encoder should not only gather information from the related utterances but also keep the information from the whole conversational context.

To avoid the information vanishing in the final conversational vector representation, we propose to use a simple but efficient reconstruction task to help the conversational encoder to keep the overall semantic information. In this task, we train the model to reconstruct the bag-of-words (BOW) vector of the whole conversation using the representation of [CLS] produced by the conversational encoder. Specifically, all of the words appearing in the context are converted to a BoW vector y^w,

$$y^w = \text{BOW}(\mathcal{S}(\text{tokenize}(Q))),$$

where the length of y^w is the vocab size and $y_i^w = 1$ only if the i-th word in vocab appears in the context otherwise $y_i^w = 0$. We use a linear layer after the last layer of the model to process $E_{[\texttt{CLS}]}$ and optimize the WR loss based on mean squared error,

$$\begin{aligned}\hat{y}^w &= \text{Sigmoid}(W_w E_{[\texttt{SEP}]} + b_w),\\ \mathcal{L}_{WR} &= \|\hat{y}^w - y^w\|_2\,,\end{aligned}$$

where $W_w \in \mathbf{R}^{h \times |V|}$, $b_w \in \mathbf{R}^{|V|}$, $|V|$ is the vocab size, $\|\cdot\|$ means euclidean distance.

4.5 Optimization

Inspired from the previous studies (Yu et al., 2021; Mao et al., 2022), we also employ the knowledge distillation objective in SSP to accelerate the learning process. Specifically, a pre-trained ad-hoc search encoder TEnc which uses the de-contextualized query as the input and produce the vector representation. We use TEnc as the teacher model and employ a knowledge distillation loss function to train our conversational encoder to mimic the vector representation produced by the teacher encoder TEnc. We formulate the knowledge distillation loss $\mathcal{L}_{KD}$ as follows:

$$\begin{aligned}E^*_{[\texttt{CLS}]} &= \text{TEnc}(\{[\texttt{CLS}], q_n^*\})_{[\texttt{CLS}]}\\ \mathcal{L}_{KD} &= \left\| E_{[\texttt{CLS}]} - E^*_{[\texttt{CLS}]} \right\|_2 .\end{aligned}$$

where the q_n^* is the manual rewritten query of q_n, $(\cdot)_{[\texttt{CLS}]}$ means only taking the [CLS] representation of TEnc's last layer output. We make the representation of conversation $E_{[\texttt{CLS}]}$ to approximate the representation of reformulation query $E^*_{[\texttt{CLS}]}$ processed by TEnc to distill its powerful retrieval ability.

Finally, we combine all the training objective of each self-supervised task and optimize all the parameters in the conversational encoder:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{KD} + \alpha\mathcal{L}_{TS} + \beta\mathcal{L}_{CI} + \gamma\mathcal{L}_{WR},$$

where the $\mathcal{L}_{\text{final}}$ is the final training objective for SSP, α, β, and γ denotes the hyper-parameter as a trade-off between the self-supervised tasks.

5 Experimental Setting

5.1 Datasets

Table 1: The statistics of test dataset for fine-tuning.

Statistics	**CAsT-19**	**CAsT-20**
# Conversations	50 (20)	25 (25)
# Queries	479 (173)	216 (208)
# Avg. Query Tokens	6.1	6.8
# Avg. Queris / Conversation	9.6	8.6
# Documents	38M	

For fine-tuning the conversational encoder on the conversational search task, we choose two few-shot datasets to evaluate our proposed model based on K-fold cross-validation.

CAsT-19 (Dalton et al., 2020) is the acronym of the TREC Conversational Assistance Track (CAsT) 2019 benchmark dataset. It is built by human annotators who are required to mimic real dialogues under specified topics and contains frequent coreferences, abbreviations, and omissions. In this work, we pay attention to query de-contextualization and but only the test set provides manual oracle de-contextualized queries. Since the queries in TREC CAsT dataset are used in the conversational search fine-tuning phrase, it will cause the data leaking problem. For a fair comparison, we filter the queries from TREC CAsT from QReCC. The statistics of the filtered QReCC dataset are shown in Table 5.

CAsT-20 (Dalton et al., 2021) refers to next year's TREC CAsT. Its most obvious modification is that the coreference could appear in the response (a summarized answer of gold passage)compared with CAsT-19, where a query only refers to its previous queries. Both manual response and automatic response (generated by neural rewriter (Yu et al., 2020)) are provided in CAsT-20. It contains 216 queries in 25 dialogues which have de-contextualized queries and most of queries have relevance judgments. Additionally, CAsT-20's corpus is the same as CAsT-19's. Detailed statistics are shown in Table 1.

5.2 Baselines

Following (Mao et al., 2022), we split baselines into two categories: sparse retrieval methods and dense retrieval methods respectively. Sparse retrieval methods rewrite the contextualized query to a context-independent query and use the ad-hoc sparse retriever to obtain the results. The dense retrieval methods use the ad-hoc dense retriever or directly encode the conversational queries via a conversational dense retriever.

• Raw denotes simply using the last context-independent query in the dense or sparse retriever to retrieve the documents.

• Tansformer++ (Vakulenko et al., 2021) is a query rewriting method which inherits from GPT-2 (Radford et al., 2019) and fine-tunes on CANARD dataset (Elgohary et al., 2019). Then it employs the ad-hoc retriever to search using the rewritten query.

• QueryRewriter (Yu et al., 2020) is a data augmentation method that first generates query reformulation data using large amounts of ad-hoc search sessions based on rules and self-supervised learning. Then the automatically generated data is used to train the query rewriter.

• QuReTeC (Voskarides et al., 2020) deals with the query reformulation task as a binary term classification problem. It will decide whether to add terms appearing in the dialogue history to the current turn query or not.

• ContQE (Lin et al., 2021a) employs a well-trained ad-hoc search encoder TCT-ColBERT (Lin et al., 2020a). It uses the mean-pooling method to get the contextual embedding and fine-tunes on pseudo-relevance labels.

• ConvDR (Yu et al., 2021) develops the few-shot learning method to train the conversational dense retriever. It takes ANCE (Xiong et al., 2020) as the teacher model to teach the conversational student model. Integrating the distilling loss and ranking loss, it obtains a pretty performance on the few-shot dataset.

• COTED (Mao et al., 2022) further introduces the curriculum denoising to inhibit the unhelpful turns in context. An additional two-step multi-task learning improves the performance of ConvDR.

• T5(WikiD+WebD) (Dai et al., 2022) trains on two large automatically generated conversational search dataset WikiDialog(11.4M dialogues) and WebDialog(8.4M dialogues) from a T5-large encoder checkpoint. Otherwise, it further warm-ups on the QReCC dataset. Though it does not fine-tune on CAsT-19 (50 dialogues) and CAsT-20 (25 dialogues), the extremely time-consuming training procedure makes its performance up to a stable level.

5.3 Evaluation Metrics

Following the previous works on conversational search, we evaluate all models based on **M**ean **R**eciprocal **R**ank (MRR) and **N**ormalized **D**iscounted **C**umulative **G**ain @3 (NDCG@3). **MRR** deems the ranking reciprocal of a positive sample as its score and counts the average of all samples. It is a simple yet effective metric for ranking tasks. **NDCG@3** considers the importance of positive samples based on their relevance and chooses scores of the top 3 samples to normalize. The statistical significance of two runs is tested using a two-tailed paired t-test and is denoted using † and ‡ for significance ($p \leq 0.05$) and strong significance ($p \leq 0.01$).

5.4 Implementation Details

Most settings in this work are similar to ConvDR (Yu et al., 2021). We employ the ad-hoc retriever ANCE (Xiong et al., 2020) as the teacher module to calculate the knowledge distillation loss. Following previous conversational search work, for CAsT-19, we concatenate the historical query and the current query as the model inputs, and we additionally take account of the historical responses for CAsT-20. The leading words in the conversational context will be truncated if the concatenation length exceeds a maximum length, which is 256 and 512 for CAsT-19 and CAsT-20 respectively. We implement experiments using PyTorch and Transformers library on an NVIDIA A40 GPU. Adam optimizer is employed with the learning rate of $2e-5$ and batch size of 64 for CAsT-19 and 32 for CAsT-20. Our model will post-train 2 epochs and then fine-tune on the conversational search corpus. The self-supervised task weights α, β and γ are set as $1e-2, 1e-3, 1e-2$ for CAsT-19 and $1e-1, 2e-3, 2e-2$ for CAsT-20. We use faiss (Johnson et al., 2019) to index the passages, whose representations are generated by ANCE and fixed. Following the TREC Conversational Assistance competition official evaluation setting, we use relevance scale ≤ 2 as positive for CAsT-19 and relevance scale ≤ 1 for CAsT-20 and obtain our result based on official evaluation scripts.

Table 2: Conversational search performance comparison. $\star$ denotes our implementation. † (‡) indicates (strong) significant improvement over ConvDR with $p \leq 0.05$ ($p \leq 0.01$).

Search	Method	CAsT-19		CAsT-20	
		MRR	NDCG@3	MRR	NDCG@3
Sparse	Raw	0.322	0.134	0.160	0.101
	Tansformer++	0.557	0.267	0.162	0.100
	QueryRewriter	0.581	0.277	0.250	0.159
	QuReTeC	0.605	0.338	0.262	0.171
Dense	Raw	0.420	0.247	0.234	0.150
	Tansformer++	0.696	0.441	0.296	0.185
	QueryRewriter	0.665	0.409	0.375	0.255
	QuReTeC	0.709	0.443	0.430	0.287
	ContQE	-	**0.499**	-	0.312
	T5(WikiD+WebD)	0.741	-	0.513	-
	COTED	0.769	0.478	0.491	0.342
	COTED*	0.758	0.475	0.481	0.321
	COTED-SSP	0.760	0.478	0.501	0.351
	ConvDR	0.740	0.466	0.501	0.340
	ConvDR-SSP	**0.780**†	0.480	**0.526**‡	**0.365**‡

6 Evaluation Result

6.1 Overall Performance

We compare our model with all baselines in Table 2. We can find that the sparse methods generally achieve less satisfying performance than the dense conversational methods, which demonstrates the dense methods can understand the search intent of users better. Our model performs consistently better on two datasets than other sparse and dense conversational search models with improvements of 1.4% and 0.4% on the CAsT-19 dataset and achieves 7.1% and 6.7% improvements on the CAsT-20 dataset compared with COTED in terms of MRR, and NDCG@3 respectively. This demonstrates that our proposed self-supervised tasks provide a useful training signal for the conversational encoder module than the simple parameter warm-up method used in previous methods.

In Table 2, we find that ContQE outperforms ConvDR-SSP on CAsT-19 in terms of NDCG@3. The possible reason is that Mao et al. (2022) has illustrated that ContQE introduces a stronger query encoder TCT-ColBERT (Lin et al., 2020a) and it takes multi-stage methods to train their conversational encoder. In contrast to the complexity of the multi-stage method, our SSP can boost the performance of the existing conversational search model in an end-to-end manner which is easier to train and deploy in real-world applications. We will leave adapting this stronger encoder TCT-ColBERT into the post-training paradigm in our future work.

To verify the generalization ability of SSP, we equip our proposed **S**elf-**S**upervised **P**ost-training to two strong conversational search methods (COTED and ConvDR), which can provide a better conversational context encoder. From the comparison between COTED and COTED-SSP, ConvDR and ConvDR-SSP, we can find that our proposed new post-train paradigm can adapt to different conversational search models and boost their performance, which demonstrate the effectiveness and generalization ability of our proposed SSP.

Table 3: Comparison between ablation models.

Method	CAsT-19		CAsT-20	
	MRR	NDCG@3	MRR	NDCG@3
ConvDR-SSP	**0.780**	**0.480**	**0.526**	**0.365**
w/o. TS	0.753	0.473	0.513	0.355
w/o. CI	0.749	0.472	0.515	0.351
w/o. WR	0.757	0.476	0.512	0.357

6.2 Ablation Study

We remove each self-supervised task to analyze the effectiveness of each component, and TS is the acronym for topic segmentation, CI denotes the coreference identification and WR denotes word reconstruction. The performance of ablation models is shown in Table 3, and we can find that all of the ablation models perform less promising than the best model ConvDR-SSP, which demonstrates the preeminence of each self-supervised task in SSP.

We ablate the topic segmentation task in ConvDR-SSP w/o. TS and observe the decline in search performance. The topic segmentation task helps the model identify the topic boundary in the long session and pay more attention to the utterances in the related topics This makes the retrieval performance raises 3.6% and 2.5% in terms of MRR on the CAsT-19 and CAsT-20 datasets respectively. In the method ConvDR-SSP w/o. CI, we remove the coreference identification self-supervised task and the performance of this ablation model dropped dramatically, which demonstrates that it plays the most important role in SSP. The experiment shows that our ConvDR-SSP achieves 4.1% and 1.7% increments compared with ConvDR-SSP w/o. CI in terms of MRR score on the CAsT-19 and CAsT-20 datasets. We also remove the word reconstruction task yielding ConvDR-SSP w/o. WR, and the dropped score shows that it is effective to keep the contextual semantic in the context representation. All of our self-supervised tasks, which provide extra supervision signals to understand dialog structure and prevent the semantic vanishing, help ConvDR-SSP achieves the best performance

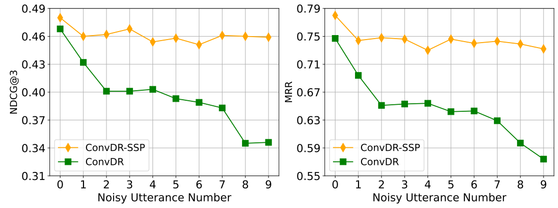

Figure 4: Robustness evaluation by adding the different numbers of off-topic utterances. We randomly sample irrelevant utterances from other search sessions and evaluate the results of ConvDR and ConvDR-SSP.

according to the experimental results.

6.3 Robustness of Topic Segmentation

To verify the effectiveness of the topic segmentation of our method, we conduct an experiment that concatenates different lengths of randomly sampled utterances to the beginning of the current conversation session. In this experiment, we use the ConvDR as our baseline. Figure 4 shows the search performance of our SSP and ConvDR with different length of random sampled noise utterances as input. From Figure 4, we find that our SSP is more robust to concatenate more random sampled utterances. When we concatenate more random sampled utterances, the performance of ConvDR dropped dramatically while ConvDR-SSP slightly dropped in the beginning and kept stable. The reason for this phenomenon lies in that our model can identify the topic segmentation boundary and reduce the impact of unrelated utterances when encoding the current conversational query. This demonstrates that the topic segmentation helps the model focus on the utterances of relevant topics.

6.4 Case Study

We show three cases in Table 4 to intuitively understand how our self-supervised tasks of SSP improve the performance of the existing conversational search methods.

In the first case, ConvDR, which equally treats every historical query, struggles with the long dialogue history and retrieves the irrelevant passage. After incorporating SSP, the topic segmentation makes ConvDR-SSP split out several most related utterances in conversational history. With the help of modeling the topic boundary, it easily discovers that “throat cancer” is the referred term for the current query.

In the second case, due to the complex historical queries, ConvDR is confused about whether the “ones” in the last query means “database” or “real-time database” and results in a unrelated retrieved passage. Our proposed coreference identification task makes ConvDR-SSP bypass these obstructions and straightly point out the referred query, and ConvDR-SSP successfully finds the accuracy result.

The contextual semantic vanishing will harm the performance since the incomplete contextual semantics cannot accurately represent the search intent. In the last case, it makes ConvDR misunderstand the meaning of “avoid” in the current query to “recover”. Then its retrieved passage mainly illustrates “how to recover from sports injuries”. The word reconstruction demonstrates its effectiveness and keeps the semantic information of “avoid”, which is indispensable during representation learning. The complete contextual semantic leads ConvDR-SSP to more accurate retrieval.

6.5 Parameter Tuning

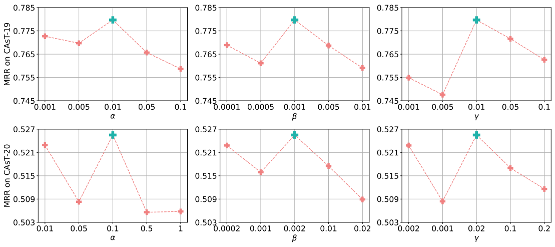

Figure 5: The parameter analysis for weights α, β and γ.

In this section, we analyze how much the hyperparameters α, β, and γ influence the retrieval performance and explore the best setting of hyperparameters. We design five-group experiments for each parameter and each dataset and the performance comparison as Figure 5. We find that the performance of ConvDR-SSP slightly drops when the parameter changes, and this demonstrates the hyper-parameter robustness of SSP. Finally, we determine the best setting of α, β, and γ to be $1e-2$, $1e-3$, $1e-2$ for CAsT-19 and $1e-1$, $2e-3$, $2e-2$ for CAsT-20.

7 Conclusion

In this work, we propose a novel **S**elf-**S**upervised **P**ost-training framework SSP for conversational search, which could easily be applied to existing methods and boost their performance. Different from the conventional warm-up method, our proposed SSP introduces three self-supervised tasks to

Table 4: Retrieved examples of ConvDR-SSP and ConvDR. We present historical queries, current query (underlined), manual reformulation query (**Ref**) and the first passages different methods disagree. The key information in the conversations and passages are marked in blue and red respectively.

Queries	**First Disagreed Passages**
CAsT Topic-31	
What is throat cancer? Is it treatable? Tell me about lung cancer. What are its symptoms? Can it spread to the throat? What causes throat cancer? What is the first sign of it? Is it the same as esophageal cancer? **Ref**:Is throat cancer the same as esophageal cancer?	ConvDR: There are two main types of esophageal cancer: squamous cell cancer and adenocarcinoma of the esophagus. Squamous cell cancer occurs most commonly in African Americans as well as people who smoke cigarettes... ConvDR-SSP: In fact, some people diagnosed with throat cancer are diagnosed with esophageal, lung, or bladder cancer at the same time. This is typically because cancers often have the same risk factors, or because cancer that begins in one part of the body can spread throughout the body...
CAsT Topic-58	
What is a real-time database? How does it differ from traditional ones? What are the advantages of real-time processing? What are examples of important ones? **Ref**:What are examples of important real-time databases?	ConvDR: Examples of what the database describes. ConvDR-SSP: A real-time database is a database systemwhich uses real-time processing to handle workloads whose state is constantly changing. This differs from traditional databases containing persistent data. For example...
CAsT Topic-59	
Which weekend sports have the most injuries? What are the most common types of injuries? What is the ACL? What is an injury for it? Tell me about the RICE method. Is there disagreement about it? What is arnica used for? What are some ways to avoid injury? **Ref**:What are some ways to avoid sports injuries?	ConvDR: To help recover from minor injuries, overexertion or surgery, Arnica is a must for every medicine cabinet. Whether you are an active baby boomer... ConvDR-SSP: Injury Prevention Basics. It's always better to prevent an injury than to recovery from one, so learning and following basic injury prevention advice is step one. The best way to avoid injuries is to be prepared for your sport, both physically and mentally. Don't succumb to the weekend warrior syndrome...

better initialize the conversational encoder. These extra supervision signals guide the model to understand complex conversational structure and effectively prevent contextual semantic vanishing. Extensive experiments conducted on two benchmark datasets prove the effectiveness of SSP, which improves previous methods and achieves the best performance. Extra analytical experiments further answer why our self-supervised tasks could improve performance.

8 Limitations

Despite we largely improve the performance of the existing conversational search method, the mechanism of the self-supervised tasks in our SSP is simple and intuitive. Additionally, our post-training method relies on the external query reformulation dataset, which is a compromise under the scarcity of conversational search data. However, the essential contribution of this work is that we point out the significance of modeling dialogue structure (especially for topic shift), and the phenomenon of contextual semantic vanishing in conversational search for the first time. We hope future works could pay more attention to these problems and devise more complex methods to develop more powerful conversational search systems.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive comments. This work was supported by National Natural Science Foundation of China (NSFC Grant No. 62122089 & No. T2293773), Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, and Intelligent Social Governance Platform, Major Innovation & Planning Inter-disciplinary Platform for the "Double-First Class" Initiative, Renmin University of China.

References

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*, pages 520–534, Online. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Zhuyun Dai and Jamie Callan. 2020. Context-aware document term weighting for ad-hoc search. In *Proceedings of The Web Conference 2020*, pages 1897–1907.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *International Conference on Machine Learning*, pages 4558–4586. PMLR.

Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. 2018. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 126–134.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2021. Cast 2020: The conversational assistance track overview. Technical report, Technical report.

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*, pages 1985–1988.

Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. 2017. Trec complex answer retrieval overview. In *TREC*.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. *Can You Unpack That? Learning to Rewrite Questions-in-Context*.

Yasuhiro Fujiwara, Makoto Nakatsuji, Hiroaki Shiokawa, Takeshi Mishima, and Makoto Onizuka. 2013. Efficient ad-hoc search for personalized pagerank. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pages 445–456.

Shen Gao, Zhaochun Ren, Yihong Zhao, Dongyan Zhao, Dawei Yin, and Rui Yan. 2019. Product-aware answer generation in e-commerce question-answering. In *WSDM*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Johannes Kiesel, Lars Meyer, Martin Potthast, and Benno Stein. 2021. Meta-information in conversational search. *ACM Transactions on Information Systems (TOIS)*, 39(4):1–44.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

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

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021a. Contextualized query embeddings for conversational search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1004–1015, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2020b. Query reformulation using query history for passage retrieval in conversational search. *arXiv preprint arXiv:2005.02230*.

Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2021b. Multi-stage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting. *ACM Transactions on Information Systems (TOIS)*, 39(4):1–29.

Kelong Mao, Zhicheng Dou, and Hongjin Qian. 2022. Curriculum contrastive context denoising for few-shot conversational dense retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 176–186.

Ida Mele, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, and Ophir Frieder. 2021. Adaptive utterance rewriting for conversational search. *Information Processing & Management*, 58(6):102682.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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*, pages 117–126.

Johanne R Trippas, Damiano Spina, Paul Thomas, Mark Sanderson, Hideo Joho, and Lawrence Cavedon. 2020. Towards a model for spoken conversational search. *Information Processing & Management*, 57(2):102162.

Quan Tu, Shen Gao, Yanran Li, Jianwei Cui, Bin Wang, and Rui Yan. 2022. Conversational recommendation via hierarchical information modeling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2201–2205, New York, NY, USA. Association for Computing Machinery.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 355–363.

Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. Query resolution for conversational search with limited supervision. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 921–930.

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. In *International Conference on Learning Representations*.

Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-shot generative conversational query rewriting. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 1933–1936.

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*, pages 829–838.

Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th acm international conference on information and knowledge management*, pages 177–186.

A Post-training Dataset

Followed by the existing conversational dense retrieval methods, we also use the query reformulation dataset for our proposed SSP model. **QReCC** (Anantha et al., 2021) is a query rewriting dataset which contains 14K conversations. The queries in QReCC are collected from three sources: TREC CAsT (Dalton et al., 2020), QuAC (Choi et al., 2018) and NQ (Kwiatkowski et al., 2019). The queries in NQ were used as prompts to create conversational queries.

We notice that the queries in TREC CAsT dataset are used in the conversational search finetune phrase, it will cause the data leaking problem. For fair comparison, we filter the queries from TREC CAsT from QReCC. The statistics of the filtered QReCC dataset are shown in Table 5.

Table 5: The statistics of QReCC after filtering queries from TREC CAsT. (Convs. means conversations, Qrs. means queies, Ref. means reformulation. '#' indicates count numbers.)

Statistics	**QReCC$_{\text{QuAC}}$**	**QReCC$_{\text{NQ}}$**
# Convs.	9124	4394
# Qrs.	62749	16455
# Avg. Convs. Tokens	162.2	87.8
# Avg. Qrs. Tokens	6.5	6.7
# Avg. Ref. Qrs. Tokens	10.4	9.1
# Avg. Qrs./Convs.	6.9	3.7

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 8

☒ A2. Did you discuss any potential risks of your work?
This work focus on conversational search, which has no obvious potential risks.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

Section 6

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Implementation Details in Appendix

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Constrained by resource, we only report resulf of the single run.

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.