External Knowledge Acquisition for End-to-End Document-Oriented Dialog Systems

Tuan M. Lai[2]*, **Giuseppe Castellucci**[1], **Saar Kuzi**[1], **Heng Ji**[1,2], **Oleg Rokhlenko**[1]
[1]Amazon
[2]University of Illinois Urbana-Champaign
{giusecas, skuzi, olegro}@amazon.com
{tuanml2, hengji}@illinois.edu

Abstract

End-to-end neural models for conversational AI often assume that a response can be generated by considering only the knowledge acquired by the model during training. Document-oriented conversational models make a similar assumption by conditioning the input on the document and assuming that any other knowledge is captured in the model's weights. However, a conversation may refer to external knowledge sources. In this work, we present EKo-DoC, an architecture for document-oriented conversations with access to external knowledge: we assume that a conversation is centered around a topic document and that external knowledge is needed to produce responses. EKo-DoC includes a dense passage retriever, a re-ranker, and a response generation model. We train the model end-to-end by using silver labels for the retrieval and re-ranking components that we automatically acquire from the attention signals of the response generation model. We demonstrate with automatic and human evaluations that incorporating external knowledge improves response generation in document-oriented conversations. Our architecture achieves new state-of-the-art results on the Wizard of Wikipedia dataset, outperforming a competitive baseline by 10.3% in Recall@1 and 7.4% in ROUGE-L.

1 Introduction

Recent advances in Natural Language Processing (NLP) allowed us to build complex conversational systems for various scenarios, such as task-oriented (Radlinski and Craswell, 2017; Wen et al., 2017), chit-chat (Khatri et al., 2018; Zhou et al., 2020), and even for guiding users in performing complex real-world tasks (Gottardi et al., 2022). When using deep neural networks for implementing a conversational agent, a common practice is to use many historical conversations to train the model to produce responses that are related to a dialog context. In this setting, a response is generated solely from the knowledge acquired during the training of the network (Vinyals and Le, 2015), and no other knowledge sources are used at inference time.

A more effective way of producing dialog responses would be to incorporate external knowledge into the model. This is, for example, the case of systems that make use of document/passage retrieval in QA (Lewis et al., 2020b). In this work, we consider the setting in which a conversation is grounded in a target topic but also in external knowledge, in the form of documents. For example, Figure 1 shows a conversation with respect to a target topic (i.e., *San Diego Comic-Con*) represented by a topic document. During the conversation, some turns may refer to other documents from an external Knowledge Base (KB) providing additional information (e.g., *Shel Dorf*). In our preliminary studies, we estimated that about 36% of the dialogs in the popular Wizard of Wikipedia dataset (Dinan et al., 2019) require knowledge beyond the topic document.

This setting poses additional challenges to the generation of adequate responses in a dialog. While a model could possibly memorize a vast amount of knowledge in its weights during training, the model will likely be applied to new dialog contexts that refer to unseen knowledge. If enriched with external knowledge, the model input could be better conditioned to produce accurate outputs. From a technical perspective, the model needs to learn i) to retrieve the relevant knowledge and ii) to incorporate it into the generated response.

To address these challenges, we propose EKo-DoC, an end-to-end conversational agent designed to model a target topic document and a set of external knowledge documents. EKo-DoC integrates a Dense Passage Retriever (DPR) (Karpukhin et al., 2020) and a Re-Ranker (RR) into a response generation model. To reduce the need

* Work done during internship at Amazon.

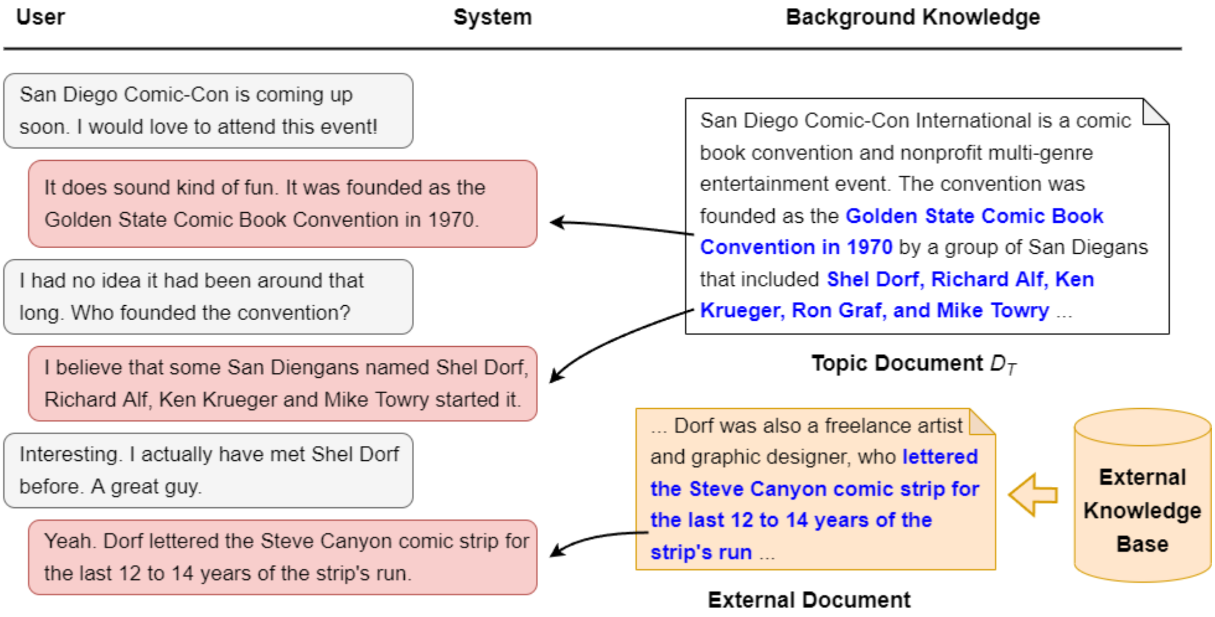

User
System
Background Knowledge
San Diego Comic-Con is coming up soon. I would love to attend this event!
It does sound kind of fun. It was founded as the Golden State Comic Book Convention in 1970.
I had no idea it had been around that long. Who founded the convention?
I believe that some San Diengans named Shel Dorf, Richard Alf, Ken Krueger and Mike Towry started it.
Interesting. I actually have met Shel Dorf before. A great guy.
Yeah. Dorf lettered the Steve Canyon comic strip for the last 12 to 14 years of the strip's run.
San Diego Comic-Con International is a comic book convention and nonprofit multi-genre entertainment event. The convention was founded as the Golden State Comic Book Convention in 1970 by a group of San Diegans that included Shel Dorf, Richard Alf, Ken Krueger, Ron Graf, and Mike Towry ...
Topic Document D_T
... Dorf was also a freelance artist and graphic designer, who lettered the Steve Canyon comic strip for the last 12 to 14 years of the strip's run ...
External Knowledge Base
External Document

Figure 1: An example of a dialog in our setting: a central topic document defines the general topic of the dialog. External documents can be used as well to generate responses for some specific turns of the conversation.

for manual annotations, we automatically acquire labels for the training of the DPR and RR components. In particular, we leverage the attention weights of the generation model over the retrieved documents to generate a set of silver labels. Our generation model is a sequence-to-sequence model that generates a response by taking as input the entire concatenation of the dialog context and related documents retrieved from a KB.

Experimental results demonstrate that EKo-DoC is able to correctly use both the topic document and the external knowledge to produce better responses in a conversation. We report competitive results on two public datasets, Wizard of Wikipedia (WoW) (Dinan et al., 2019) and MultiDoc2Dial (Feng et al., 2021). In particular, EKo-DoC achieves new state-of-the-art (SOTA) results on WoW, outperforming a strong baseline by 10.3% in Recall@1 and 7.4% in ROUGE-L. Finally, we conduct a human study to verify the quality of the generated responses. Compared to models that do not condition on the topic document or external knowledge, our model produces responses that are more fluent, more on-topic, and more interesting.

To summarize, our contributions are: i) we study the setting where we model *both* a target topic document and external knowledge for a conversation; ii) we propose EKo-DoC, a novel end-to-end response generation architecture for the task; iii) we propose an automatic annotation procedure to acquire the labels for training the retrieval engine.

2 EKo-DoC Architecture

Given a dialog context C and a topic document D_T that serves as the background of the dialog, the task is to generate a response r to the last utterance of C. We also have available an external knowledge base $K_{\text{KB}} = \{D_1, D_2, ..., D_m\}$; D_j denotes a knowledge snippet represented by some natural language text (e.g., a paragraph in Wikipedia). While we use the term “snippet”, each snippet is not necessarily a short piece of text. EKo-DoC is general enough to be used with documents of different lengths.

The dialog is mostly centered around D_T, but external knowledge from K_{KB} is sometimes needed to generate an informative and relevant response (see Figure 1 for an example). To this end, the task can be viewed as building a model of $P(r|C, D_T, K_{\text{KB}})$.

2.1 Document-Oriented Dialog System with Access to External Knowledge

Figure 2 shows an overview of EKo-DoC, our proposed framework. The inference process consists of three steps. First, given a dialog context and a topic document, we use a DPR model to retrieve an initial set of knowledge snippets (Section 2.1.1). Second, we re-rank the snippets using a Transformer-based cross-attention model (Section 2.1.2). Third, conditioned on the dialog context and

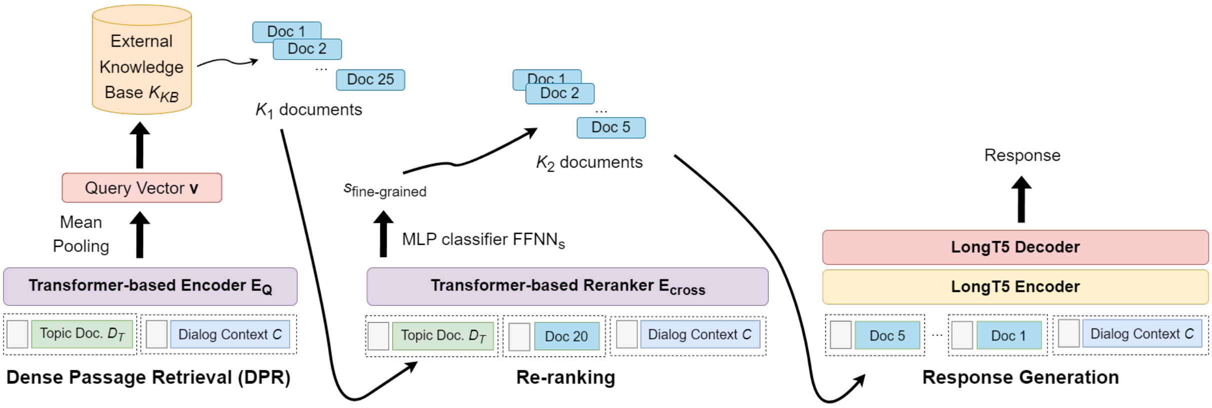


Figure 2: An overview of EKo-DoC, our framework. K_1 and K_2 are empirically set to be 25 and 5 (respectively).

the top-k ranked snippets, a generator produces a natural language response (Section 2.1.3).

In previous work (Glass et al., 2022; Li et al., 2022a), the retrieval components were trained using human-annotated pairs of input contexts and supporting knowledge snippets. To remove the need for expensive gold-standard retrieval labels, we use attention scores as pseudo-labels for training the retrieval and re-ranking models (Section 2.2).

Finally, note that while D_T can be fundamental to the response generation for most of the conversation, it can be less relevant for some dialog turns (e.g., see Figure 1). Therefore, we also include D_T in the knowledge base K_{KB} and let our retrieval engine decide if the topic document is essential. If D_T is retrieved, the final generator will use D_T for response generation.

2.1.1 External Knowledge Retrieval

In order to use a DPR model for retrieving knowledge, we need to encode K_{KB} in the indexing phase and the input query in the retrieval phase.

K_{KB} Encoding. We use an encoder $E_{\text{KB}}(\cdot)$ to map every knowledge snippet D_j to a real-valued vector. We assume that each snippet is a document with a title and short textual content. The input to the document encoder $E_{\text{KB}}(\cdot)$ is:

$$T\Big[D\Big] = \texttt{[s]}\ \text{Title / Description}\ \texttt{[/s]} \qquad (1)$$

where [s] and [/s] are special tokens denoting the start and end of the input, and "/" is used to separate the title from the description. $T[\cdot]$ denotes a function that maps a general object (e.g., a knowledge snippet) into a textual sequence.

We use a pre-trained Transformer model as the encoder (Reimers and Gurevych, 2019). That is, for each $D_j \in K_{\text{KB}}$, we encode it into a vector $\mathbf{v}_j$:

$$\mathbf{v}_j = E_{\text{KB}}\Big(T\Big[D_j\Big]\Big) \qquad (2)$$

Query Encoding. We use the topic document D_{T} and the dialog context C to construct a query for the DPR model. We first concatenate D_{T} and C into a single sequence $T\Big[D_{\text{T}} \oplus C\Big]$:

$$\texttt{[s]}\ T\Big[D_{\text{T}}\Big]\ \texttt{[sep]}\ T\Big[C\Big]\ \texttt{[/s]} \qquad (3)$$

where [sep] is a special token to separate the topic document and the dialog context. $T\Big[D_{\text{T}}\Big]$ is similar to what described in Equation 1. $T\Big[C\Big]$ is instead a concatenation of all the dialog turns, where two turns are separated by the [sep] token.

Then, we also use a Transformer encoder $E_{\text{Q}}(\cdot)$ to map C and D_{T} into a single vector **v**:

$$\mathbf{v} = E_{\text{Q}}\Big(T\Big[D_{\text{T}} \oplus C\Big]\Big) \qquad (4)$$

In this work, $E_Q(\cdot)$ and $E_{KB}(\cdot)$ share the same architecture (Reimers and Gurevych, 2019).

Scoring and Retrieval. The first (coarse-grained) relevance score of a knowledge snippet D_j with respect to a query is given by:

$$s_{\text{c}}\Big(D_j, C, D_{\text{T}}\Big) = \mathbf{v} \cdot \mathbf{v}_j \qquad (5)$$

We retrieve up to K_1 knowledge snippets by ranking according to the s_c values. K_1 is a hyper-parameter, with $K_1 \ll |K_{\text{KB}}|$. In order to make this operation efficient, we adopted FAISS (Johnson et al., 2021). We first apply $E_{\text{KB}}(\cdot)$ to all the knowledge snippets and index them using FAISS offline. Then, given a query vector **v** obtained with $E_Q(\cdot)$, we use FAISS to return the top K_1 candidates according to the coarse-grained scores.

2.1.2 Knowledge Snippet Re-ranking

After the initial coarse-grained retrieval step with DPR, there are K_1 candidate knowledge snippets. We then apply a fine-grained cross-attention re-ranker over the K_1 snippets. After that, we keep the top-K_2 snippets[1] for the next phase (i.e., the response generation phase).

The input to the re-ranker is the concatenation of D_{T}, a knowledge snippet D_j retrieved in the previous phase, and C:

$$\texttt{[s]}\ T\big[D_{\mathrm{T}}\big]\ \texttt{[sep]}\ T\big[D_j\big]\ \texttt{[sep]}\ T\big[C\big]\ \texttt{[/s]} \tag{6}$$

The re-ranker consists of a Transformer-based encoder and a feed-forward neural network. Given an input representation described above, the re-ranker computes the final relevance score s_f as:

$$\begin{aligned} \mathbf{h}_j &= \mathrm{reduce}\Big(E_{\text{cross}}\Big(T\big[D_{\mathrm{T}} \oplus D_j \oplus C\big]\Big)\Big) \\ s_{\mathrm{f}}\Big(D_j, C, D_{\mathrm{T}}\Big) &= \mathrm{FFNN}_s\big(\mathbf{h}_j\big) \end{aligned} \tag{7}$$

where $E_{\text{cross}}(\cdot)$ is a Transformer-based encoder (Liu et al., 2019), and its input is a representation described in Equation 6; FFNN_s is a feed-forward neural network. $\mathrm{reduce}(\cdot)$ is a function that returns the final hidden state of the Transformer that corresponds to the first input token. The final set of K_2 knowledge snippets is selected according to the s_{f} scores assigned from the re-ranker.

2.1.3 Response Generation Model

Our generative model is based on the sequence-to-sequence (seq2seq) encoder-decoder architecture that directly predicts an output sequence from an input sequence (Sutskever et al., 2014). By leveraging the recent advances in neural models for long sequences (Beltagy et al., 2020; Guo et al., 2022), our generator takes as input the entire concatenation of the dialog context and the top-ranked retrieved knowledge snippets (Figure 2). We use LongT5 as our generator (Guo et al., 2022) as it can scale up to 16K input length.

Specifically, let $\{\tilde{D}_1, \tilde{D}_2, ..., \tilde{D}_{K_2}\}$ be the set of knowledge snippets retrieved by the re-ranker.[2] The input to the generative model is:

$$\texttt{[s]}\ T\big[\tilde{D}_1 \oplus \tilde{D}_2 \oplus ... \oplus \tilde{D}_{K_2}\big]\ \texttt{[sep]}\ T\big[C\big]\ \texttt{[/s]} \tag{8}$$

where $T\big[\tilde{D}_1 \oplus \tilde{D}_2 \oplus ... \oplus \tilde{D}_{K_2}\big]$ is the concatenation of all the retrieved snippets:

$$T\big[\tilde{D}_1\big]\ \texttt{[sep]}\,T\big[\tilde{D}_2\big]\ \texttt{[sep]}...\ \texttt{[sep]}\,T\big[\tilde{D}_{K_2}\big] \tag{9}$$

We denote the final input sequence as X. The expected output is an appropriate response Y to the last dialog turn. Similar to previous encoder-decoder language models (Lewis et al., 2020a; Raffel et al., 2020), the generator models the conditional probability of selecting a new token given all previously generated tokens when conditioned on X:

$$P(Y|X) = \prod_{i=1}^{|Y|} P\big(Y_i \,|\, Y_{1:i-1}\,, X\big) \tag{10}$$

where Y_i is the i-th token of Y, and $Y_{1:i-1}$ consists of all the tokens that come before it.

The loss function for training the generator is the usual negative log-likelihood function:

$$\mathcal{L}_{\text{generation}}(B) = -\sum_{i=1}^{|B|} \log P\big(\hat{Y}^i | X^i\big) \tag{11}$$

where B is a mini-batch consisting of $|B|$ examples, each in the form $\big(X^i, \hat{Y}^i\big)$. In addition, $\hat{Y}^i$ is the ground-truth response to the input X^i.

We refer to our approach as the Fusion-in-Input (FiI) method, as our generator directly takes as input all the potentially relevant information (i.e., the retrieved knowledge and the dialog history). This is a departure from the popular Fusion-in-Decoder (FiD) approach that encodes the retrieved snippets independently (Izacard and Grave, 2021a,b; Asai et al., 2022).

2.2 Using Cross-Attention Scores as Retrieval Pseudo-Labels

If we train our system using only the generation loss function $\mathcal{L}_{\text{generation}}(B)$, only the parameters of the generation model will be updated. In order to optimize the retrieval engine, we propose to use the generation attention scores over the snippets in the input as retrieval pseudo-labels during the training stage. The intuition is that if the decoder of the generator pays more attention to a particular retrieved knowledge snippet, it means that such a snippet is likely to be relevant.

For each snippet retrieved by the re-ranker, we average all the pre-attention scores over all the tokens of the snippet. Formally, let L denote the number of layers of the decoder, and let H be the

[1] Note that $K_2 < K_1$.

[2] Note that it is possible that D_T is an element of this set.

number of attention heads. For each knowledge snippet D_j, we denote the indices of its starting and ending tokens in X as s_j and e_j, respectively. The averaged attention score $\mathcal{A}(D_j)$ of D_j is computed as:

$$\mathcal{A}(D_j) = \frac{\sum_{l=1}^{L} \sum_{h=1}^{H} \sum_{x=s_j}^{e_j} \sum_{y=1}^{|\hat{Y}|} p_{l,h}(x, y)}{L \times H \times |s_j - e_j + 1| \times |\hat{Y}|} \tag{12}$$

where $\hat{Y}$ is the ground-truth response. In addition, $p_{l,h}(x, y)$ is the pre-attention score between the y-th token of $\hat{Y}$ and the x-th token of X computed by the h-th attention head of the l-th layer. Basically, we collect all the pre-attention scores that are relevant to D_j and compute their average.

We can normalize the averaged attention scores using a softmax function:

$$P_{\text{attention}}(D_j) = \frac{\exp\left(\mathcal{A}(D_j)\right)}{\sum_{i=1}^{K_2} \exp\left(\mathcal{A}(D_i)\right)} \tag{13}$$

Similarly, we can normalize the retrieval scores predicted by the DPR model and the reranker:

$$\begin{aligned} Q_{\text{coarse}}(D_j) &= \frac{\exp\left(s_{\text{c}}(D_j, C, D_T)\right)}{\sum_{i=1}^{K_2} \exp\left(s_{\text{c}}(D_i, C, D_T)\right)} \\ Q_{\text{fine}}(D_j) &= \frac{\exp\left(s_{\text{f}}(D_j, C, D_T)\right)}{\sum_{i=1}^{K_2} \exp\left(s_{\text{f}}(D_i, C, D_T)\right)} \end{aligned} \tag{14}$$

We then use the averaged attention scores of the retrieved knowledge snippets as pseudo-labels to update the retrieval engine:

$$\begin{aligned} \mathcal{L}_{\text{dpr}}(B) &= \sum_{i=1}^{|B|} D_{\text{KL}}\left(P^i_{\text{attention}} \mid\mid Q^i_{\text{coarse}}\right) \\ \mathcal{L}_{\text{reranker}}(B) &= \sum_{i=1}^{|B|} D_{\text{KL}}\left(P^i_{\text{attention}} \mid\mid Q^i_{\text{fine}}\right) \\ \mathcal{L}_{\text{retrieval}}(B) &= \mathcal{L}_{\text{dpr}}(B) + \mathcal{L}_{\text{reranker}}(B) \end{aligned} \tag{15}$$

where B is a mini-batch of examples, and D_{KL} denotes the KL divergence. Also, $P^i_{\text{attention}}$, Q^i_{coarse}, and Q^i_{fine} are the distributions computed for the i-th example of B (refer to Equations 13 and 14).

The final loss function combines the generation loss (Eq. 11) and the retrieval loss (Eq. 15):

$$\mathcal{L}_{\text{final}}(B) = \frac{1}{|B|}\left(\mathcal{L}_{\text{generation}}(B) + \mathcal{L}_{\text{retrieval}}(B)\right)$$

A related study (Izacard and Grave, 2021a) also proposed an approach to train retrieval systems without strong supervision. The work focuses on the task of question answering and uses the FiD approach for output generation. In contrast, we focus on building dialog systems and tailoring our learning approach to our newly proposed FiI approach.

To summarize, EKo-DoC is end-to-end trainable. The document encoder of the DPR model (Equation 1) is fixed. The query encoder (Equation 4), the re-ranker, and the generator are optimized end-to-end using the final loss function described above.

3 Experiments

3.1 Data and Experiments Setup

Datasets For the experiments, we use the following two publicly available conversational datasets.

Wizard of Wikipedia (WoW) is a dataset consisting of open-domain conversations grounded in knowledge from Wikipedia (Dinan et al., 2019). We use the version provided by the KILT benchmark (Petroni et al., 2021). The external knowledge base K_{KB} consists of about 22 million 100-word passages from Wikipedia. We consider each passage as a knowledge snippet. Each conversation in WoW is annotated with a topic, which is a Wikipedia page. As the topic Wikipedia page can be extremely long, we only use the first paragraph of the page as the topic document D_T. Finally, WoW already comes with a public train/dev/test split.

MultiDoc2Dial is a new conversational dataset that grounds dialogs in multiple documents (Feng et al., 2021). Each dialog consists of multiple segments, and two adjacent segments are grounded in different documents. K_{KB} consists of 4,283 passages across four domains (Social Security Administration, Veteran Affairs, Student-Aid, and DMV). For each dialog, we use the first paragraph of the first grounded document as D_T.

Note that the original forms of WoW and MultiDoc2Dial do not have any notion of a central topic document. Therefore, we use some heuristics (described above) to select the topic document.

Evaluation Metrics For retrieval, we compute Recall@k, which measures the fraction of times the correct knowledge snippet is found in the top-k retrieved snippets. We evaluate the text generation output based on unigram F1 score (F1) and ROUGE-L (Lin, 2004). For MultiDoc2Dial, similarly to the original study that first introduced it

(Feng et al., 2021), we also compute Exact Match (EM) (Rajpurkar et al., 2016) and SacreBLEU (BL) (Post, 2018). For WoW, we do not report EM and BL scores because the leaderboard for this dataset does not use these metrics, and we also do not have direct access to the test set used by the leaderboard.

Baselines We implemented several baselines for detailed comparison and analysis. The first set of baselines consists of basic seq2seq models that do not have any retrieval engine:

- LongT5 (Guo et al., 2022) is an extension of T5 (Raffel et al., 2020) that handles long sequence inputs efficiently. The baseline takes only the dialog context C as input.
- LongT5-with-Topic uses the same architecture as LongT5 but takes both the topic document D_T and the dialog context C as input.

The second set of baselines consists of systems augmented with *frozen* retrievers:

- [Frozen DPR + FiI] first uses a pre-trained DPR model to retrieve five potentially relevant knowledge snippets. Conditioned on the retrieved knowledge and the dialog context, a FiI generation model then generates a response.
- [Frozen DPR + FiI (using D_T)] uses the same architecture, but its FiI generator also takes the topic document as input.

By comparing our full model against this set of baselines, we can analyze the effectiveness of using cross-attention scores as retrieval pseudo-labels.

Hyperparameters We initialized the DPR component using `all-mpnet-base-v2` (Reimers and Gurevych, 2019). The reranker is initialized using `distilroberta-base`, a distilled version of RoBERTa (Liu et al., 2019). We initialize the generation component of every system using a large version of LongT5 (Guo et al., 2022) unless otherwise stated. The batch size is 32. The number of training epochs is 10. More details about the hyperparameters are in the appendix.

3.2 Experimental Results

Table 1 presents the retrieval and response generation performance of our models and the baselines. First, we can see that our models achieve the highest performance compared with all baselines for both data sets and all evaluation metrics. This result emphasizes the importance of using external knowledge for document-oriented dialog systems and demonstrates the effectiveness of doing so by using our approach of fine-tuning the retrieval engine with attention score-based pseudo-labels.

The results in Table 1 also show that using the topic document is crucial for effective response generation. For example, LongT5-with-Topic outperforms the default LongT5 baseline, especially on MultiDoc2Dial. Similar results are observed for the baselines that use frozen retrievers.

The results also show that there is a clear positive correlation between text generation performance and retrieval performance. Compared with the frozen retrieval baselines, our full model achieves much higher generation performance by training the retrieval and re-ranking models with our end-to-end approach. The generator of every model listed in Table 1 is initialized with the same large version of LongT5. This shows the importance of the retrieval engine on the response generation performance.

Using cross-attention scores as retrieval pseudo-labels is highly effective according to Table 1. Specifically, `EKo-DoC` achieves a Recall@1 score of 71.21% on WoW without using human-annotated retrieval labels. Notice also the importance of the re-ranker component. When the re-ranker is disabled (second row) the model achieves a Recall@1 score of 33.88% on MultiDoc2Dial, while the full model achieves a score of 41.60%. Furthermore, Table 3 shows that the effectiveness of using cross-attention scores can even be comparable to that of using gold retrieval labels.

Finally, many previous studies on knowledge-intensive dialog systems do not have any notion of a central topic document (Shuster et al., 2021; Paranjape et al., 2022). To directly compare against these studies, we also train a system that does not explicitly use the topic document (Table 2). Basically, this system is similar to our full model (Figure 2), but it does not take the topic document as part of its input. Overall, our system outperforms all previous state-of-the-art methods in most evaluation metrics. While FiD-Light (Hofstätter et al., 2022) and Re2G (Glass et al., 2022) achieve better Recall@1 and Recall@5 than ours (respectively), FiD-Light and Re2G use gold retrieval labels to train their retrieval engines. Our system does not use such supervision signals, which can be expen-

	Wizard of Wikipedia			MultiDoc2Dial				
	F1	ROUGE-L	Recall@1	F1	ROUGE-L	Recall@1	EM	BL
Our Models								
EKo-DoC *	**20.77**	**18.64**	**71.21**	**42.80**	**39.02**	**41.60**	**6.40**	**29.17**
EKo-DoC (No Re-ranker) *	20.72	18.63	69.71	41.31	37.47	33.88	6.13	27.36
Baselines with Frozen Retrievers								
Frozen DPR + FiI (using D_T) *	19.66	17.80	49.69	36.19	32.44	14.44	4.20	21.30
Frozen DPR + FiI	19.29	17.53	49.69	33.90	30.31	14.44	3.74	18.56
Seq2Seq Baselines								
LongT5-with-Topic *	19.92	18.12	-	30.12	26.43	-	2.27	13.00
LongT5	16.21	14.94	-	24.47	20.68	-	0.66	5.46

Table 1: Overall results (in %) on the test sets of WoW and MultiDoc2Dial. For fair comparison, all models shown here do not use any gold retrieval labels during training. The symbol * denotes models that explicitly use the central document. In such a model, at least one component directly includes the central document as part of its input. All differences in performance between our models and the baselines are statistically significant with a p-value < 0.05.

	Recall@1	Recall@5	ROUGE-L	F1	KILT-F1
EKo-DoC (**Without** explicitly using the central doc.)	61.86	78.18	**18.32**	**20.42**	**14.41**
QKConv (Cai et al., 2022)	60.98	76.58	17.72	19.95	13.64
Hindsight (Paranjape et al., 2022)	56.08	74.27	17.06	19.19	13.39
FiD-Light (Hofstätter et al., 2022)	**66.15**	76.51	15.78	17.82	13.06
Re2G (Glass et al., 2022)	60.10	**79.98**	16.76	18.90	12.98
SEAL (Bevilacqua et al., 2022)	57.55	78.96	16.65	18.34	11.63
RAG (Petroni et al., 2021)	57.75	74.61	11.57	13.11	8.75

Table 2: Performance of state-of-the-art models on the test set of WoW according to the public leaderboard (as of February 2023). We are hiding our score on the leaderboard during the anonymous review process. The leaderboard is available at https://eval.ai/web/challenges/challenge-page/689/leaderboard/1909, and it also uses additional metrics such as Recall@5 and KILT-F1 (Petroni et al., 2021). Note that FiD-Light and Re2G use gold retrieval labels to train their retrieval engines, while our system does not rely on such supervision signals.

	R@1	R@5
EKo-DoC	45.32	61.36
DPR + RR (finetuned using gold labels)	**46.46**	**65.88**

Table 3: Comparison between using attention score-based pseudo-labels and using gold retrieval labels. Here, scores on the dev set of WoW are reported. In addition, different from the public leaderboard, we evaluate retrieval at a more fine-grained granularity, which is the passage level instead of the page level.

sive to obtain.

3.3 Human Evaluation

There can be many appropriate responses to the last turn of a dialog; therefore, human evaluation is typically crucial to properly evaluate the performance of a dialog system (Liu et al., 2016; Gandeharioun et al., 2019). We conducted a human evaluation of various models by using Amazon's Mechanical Turk (AMT).[3] Specifically, we created 150 evaluation tasks, each consisting of a dialog context selected randomly from WoW and a set of responses produced by different model variants. For each task, we asked three different AMT workers to rank the models' outputs based on:

1. *Fluency* (0/1). Is the generated response fluent and grammatically correct?
2. *Relevance* (0/1). Is the response on-topic and relevant to the last turn of the dialog?
3. *Interestingness* (0/1). Does the response provide new interesting information that is not already mentioned in the dialog?

Table 4 presents the human evaluation results comparing our full model against two baselines, LongT5 and LongT5-with-Topic. We chose these baselines as our main goal here is to analyze how external knowledge helps in document-oriented dialogs. According to human annotators, our model

[3] All details about the annotation tasks are in the appendix.

outperforms both baselines substantially in all three criteria. Furthermore, EKo-DoC achieves the most gain in the "interestingness" criterion. This is expected because, for example, compared to LongT5-with-Topic, our model also makes use of external knowledge when generating responses.

Full Model vs. LongT5	Better	Same	Worse
Fluency	31.01	47.29	21.71
Relevance	39.85	38.35	21.80
Interestingness	48.39	35.48	16.13
Full Model vs. LongT5-with-Topic			
Fluency	24.62	62.31	13.08
Relevance	32.03	53.13	14.84
Interestingness	40.60	39.85	19.55

Table 4: Human evaluation results (%) comparing our full model to LongT5 (top part) and LongT5-with-Topic (bottom part). The numbers in the column "Better" denote the percentage of the snippets where our full model is considered better (according to humans).

3.4 Analysis

Qualitative Analysis Table 5 shows some examples from WoW that illustrate how incorporating external knowledge can improve the quality of generated responses. Note that more qualitative examples are provided in the appendix.

In the first example, the responses generated by all models are fluent and grammatically correct. However, LongT5 generates a factually incorrect statement, as it does not have any knowledge that is not implicitly stored in its parameters. Nevertheless, LongT5 seems to possess some knowledge about the world, as Sam Walton is the founder of Walmart, which is related to Target Corporation.

In the second example, in the beginning, the dialog is mainly about rock and roll in general. However, the last dialog turn deviates slightly from the general topic and starts talking about James Brown, who influenced several music genres. The topic document does not have any information about the person. Therefore, LongT5-with-Topic generates a factually correct response about rock and roll, but it is unrelated to James Brown. In contrast, by leveraging external knowledge, our model generates a response that is more relevant to the last dialog turn.

Comparison between FiI and FiD We compare our FiI method and the popular FiD method (Izacard and Grave, 2021b) in Table 6. FiI outperforms FiD, as FiI encodes the dialog context and the top-ranked knowledge snippets jointly. In contrast, FiD encodes the retrieved snippets independently, so there is less interaction between the knowledge snippets before the final decoding stage. In conclusion, by using FiI instead of FiD, our generator can extract semantic relationships between the different input signals more effectively.

Examples

Topic Document: The first paragraph of the Wikipedia page talking about Target Corporation.

Dialog Context:

A: Target Corporation is the second largest discount retailer.

B: I assume that it is just before Walmart, but still huge.

Candidate Responses:

LongT5: It was founded in 1971 by Sam Walton.

LongT5-with-Topic: Target was founded by George Dayton in 1902.

EKo-DoC: Target is the eighth largest retailer in the US.

Topic Document: The first paragraph of the Wikipedia page talking about Rock and roll.

Dialog Context:

A: My favorite music is rock and roll.

B: Gotta love some Rock and Roll. It's one of my favorites as well.

...

A: Indeed it did. I was speaking to a friend earlier about *James Brown* and how he influenced so many genres of music starting in the 60's.

Candidate Responses:

LongT5-with-Topic: I like the fact that while elements of what was to become rock and roll can be heard in blues records from the 1920s and in country records of the 1930s, the genre did not acquire its name until 1954.

EKo-DoC: I agree. He was an American singer, songwriter, dancer, musician, record producer and bandleader.

Table 5: Examples showing how external knowledge improves the quality of generated responses. Red is used to indicate responses that are factually *incorrect*. Blue is used to indicate responses that are factually *correct*.

	F1	ROUGE-L
Freezed DPR + FiI (LongT5)	**34.07**	**29.97**
Freezed DPR + FiD (LongT5)	33.82	29.77
Freezed DPR + FiD (T5)	32.80	28.65

Table 6: Comparison between FiI and FiD. The dev set of MultiDoc2Dial is used.

4 Related Work

Document-oriented Conversations Humans typically seek information in a *conversational* manner, for example to find answers to questions (Choi et al., 2018) or to seek guidance in performing real world tasks (Gottardi et al., 2022). As such, several recent conversational datasets were introduced for building models to assist in information-seeking dialogs. For example, Choi et al. (2018) presented QuAC, a dataset that contains 14K information-seeking QA dialogs:

each is centered around a short evidence text from Wikipedia and involves a student and a teacher. The student poses a sequence of free-form questions about the text, while the teacher answers the questions by providing short excerpts from the text. Concurrently, Reddy et al. (2019) introduced CoQA, a dataset in which a machine has to understand a text passage and answer a series of questions that appear in a conversation. Choi et al. (2022) introduced a dataset for the novel setting of Conversational Task Assistants, where users seek guidance from a conversational agent to perform real world tasks: in this setting, a conversation is centered around a document describing a task. SOTA methods for these tasks typically do not use any external knowledge beyond the dialog context and the given background text (Zhu et al., 2018; Huang et al., 2019; Qu et al., 2019; Gupta et al., 2020; Zhao et al., 2021; Qian et al., 2022).

Open-Domain Question Answering In open-domain QA, the goal is to find the answer to a (typically short) question over a large corpus such as Wikipedia (Voorhees and Tice, 2000; Chen and Yih, 2020). Passage retrieval has been an essential component of many state-of-the-art open-domain QA systems (Karpukhin et al., 2020; Lewis et al., 2020b; Piktus et al., 2021; Min et al., 2021; Zhu et al., 2021). An effective retrieval component can reduce the search space for answer extraction and identify the support context for users to verify the answer. As a result, many studies have focused on improving the retrieval components, ranging from removing the need for strong supervision signals (Izacard and Grave, 2021a; Ram et al., 2022) to adding sophisticated reranking components (Yu et al., 2022a; Glass et al., 2022; Yu et al., 2022b). Open-domain QA is typically non-conversational and does not have any notion of a topic document.

Weak Supervision for Neural Retrieval A closely related study (Izacard and Grave, 2021a) also proposed an approach to train retrieval systems without strong supervision. The work focuses on the task of QA and uses the FiD approach for output generation. In contrast, we focus on building dialog systems and tailoring our learning approach to our newly proposed FiI approach.

Knowledge-Grounded Dialog Generation Incorporating background knowledge into conversation models can make dialogs more informative and engaging. Therefore, many recent studies have investigated various techniques for selecting relevant knowledge and integrating it into the response generation process (Lian et al., 2019; Kim et al., 2020; Li et al., 2020; Shuster et al., 2021; Chen et al., 2021; Mishra et al., 2022; Li et al., 2022b). Many of these studies utilize the Wizard of Wikipedia (WoW) dataset (Dinan et al., 2019). Our system, `EKo-DoC`, achieves a new state-of-the-art result on WoW, as shown in Table 2. This result provides some guarantee on the relative performance of `EKo-DoC` compared to many previous SOTA knowledge-grounded dialog systems. As of February 2023, there are nearly 40 tested methods on the public leaderboard of WoW, though Table 2 only shows a subset due to space limitations.

5 Conclusions

This work proposes and studies a new problem setting that combines document-oriented conversations and open-domain QA. We introduce a new architecture for the problem that includes a dense passage retriever, a re-ranker, and a response generation model. We train these three components end-to-end and use cross-attention scores as pseudo-labels to update the retrieval engine. Extensive experimental results on two public datasets demonstrate the effectiveness of our method. In the future, we plan to reduce the computational complexity of our model by using compression techniques.

6 Ethical Consideration

Limitations While `EKo-DoC` achieves new SOTA results on WoW, its performance is far from perfect. A ROUGE-L score of 18.64% and a Recall@1 score of 71.21% indicate that there is much room for improvement. Based on our manual analysis, we found that `EKo-DoC` sometimes generates responses that are a little bit unnatural. Finally, a common limitation of many Transformer-based systems, such as `EKo-DoC`, is the large computational complexity. We plan to reduce the computational complexity of `EKo-DoC` by using some compression techniques (Lai et al., 2020; Sun et al., 2020).

Potential Risks A potential malicious use case of research on conversational AI is for building dialog systems that pose as humans and then proactively alter users' perceptions about specific issues, evaluations of products or services, or political inclinations (Qi et al., 2021). We urge anyone who uses or builds upon our research to avoid such malicious use cases.

References

Akari Asai, Matt Gardner, and Hannaneh Hajishirzi. 2022. Evidentiality-guided generation for knowledge-intensive NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2226–2243, Seattle, United States. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *ArXiv*, abs/2004.05150.

Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Wen-tau Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive search engines: Generating substrings as document identifiers. *arXiv preprint arXiv:2204.10628*.

Mingzhu Cai, Siqi Bao, Xin Tian, H. He, Fan Wang, and Hua Wu. 2022. Query enhanced knowledge-intensive conversation via unsupervised joint modeling. *ArXiv*, abs/2212.09588.

Danqi Chen and Wen-tau Yih. 2020. Open-domain question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 34–37, Online. Association for Computational Linguistics.

Xiuyi Chen, Feilong Chen, Fandong Meng, Peng Li, and Jie Zhou. 2021. Unsupervised knowledge selection for dialogue generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1230–1244, 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.

Jason Ingyu Choi, Saar Kuzi, Nikhita Vedula, Jie Zhao, Giuseppe Castellucci, Marcus Collins, Shervin Malmasi, Oleg Rokhlenko, and Eugene Agichtein. 2022. Wizard of tasks: A novel conversational dataset for solving real-world tasks in conversational settings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3514–3529, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah J. Jones, Àgata Lapedriza, and Rosalind W. Picard. 2019. Approximating interactive human evaluation with self-play for open-domain dialog systems. In *NeurIPS*.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Rajaram Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2g: Retrieve, rerank, generate. *arXiv preprint arXiv:2207.06300*.

Anna Gottardi, Osman Ipek, Giuseppe Castellucci, Shui Hu, Lavina Vaz, Yao Lu, Anju Khatri, Anjali Chadha, Desheng Zhang, Sattvik Sahai, Prerna Dwivedi, Hangjie Shi, Lucy Hu, Andy Huang, Luke Dai, Bofei Yang, Varun Somani, Pankaj Rajan, Ron Rezac, Michael Johnston, Savanna Stiff, Leslie Ball, David Carmel, Yang Liu, Dilek Hakkani-Tur, Oleg Rokhlenko, Kate Bland, Eugene Agichtein, Reza Ghanadan, and Yoelle Maarek. 2022. Alexa, let's work together: Introducing the first alexa prize taskbot challenge on conversational task assistance. In *Alexa Prize TaskBot Challenge Proceedings*.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Somil Gupta, Bhanu Pratap Singh Rawat, and Hong Yu. 2020. Conversational machine comprehension: a literature review. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2739–2753, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sebastian Hofstätter, Jiecao Chen, Karthik Raman, and Hamed Zamani. 2022. Fid-light: Efficient and effective retrieval-augmented text generation. *ArXiv*, abs/2209.14290.

Hsin-Yuan Huang, Eunsol Choi, and Wen-tau Yih. 2019. Flowqa: Grasping flow in history for conversational machine comprehension. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Gautier Izacard and Edouard Grave. 2021a. Distilling knowledge from reader to retriever for question answering. In *International Conference on Learning Representations*.

Gautier Izacard and Edouard Grave. 2021b. Leveraging passage retrieval with generative models for open domain question answering. In *EACL*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

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

Chandra Khatri, Behnam Hedayatnia, Chandra Khatri, Anushree Venkatesh, Jeff Nunn, Yi Pan, Qing Liu, Han Song, Anna Gottardi, Sanjeev Kwatra, Sanju Pancholi, Ming Cheng, Qinglang Chen, Lauren Stubel, and Karthik Gopalakrishnan. 2018. Advancing the state of the art in open domain dialog systems through the alexa prize. In *Alexa Prize SocialBot Grand Challenge 2 Proceedings*.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential Latent Knowledge Selection for Knowledge-Grounded Dialogue. In *ICLR*.

Tuan Manh Lai, Quan Hung Tran, Trung Bui, and Daisuke Kihara. 2020. A simple but effective bert model for dialog state tracking on resource-limited systems. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8034–8038. IEEE.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Kun Li, Tianhua Zhang, Liping Tang, Junan Li, Hongyuan Lu, Xixin Wu, and Helen Meng. 2022a. Grounded dialogue generation with cross-encoding re-ranker, grounding span prediction, and passage dropout. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 123–129, Dublin, Ireland. Association for Computational Linguistics.

Linxiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-resource knowledge-grounded dialogue generation. *Advances in Neural Information Processing Systems*, 33:8475–8485.

Sha Li, Mahdi Namazifar, Di Jin, Mohit Bansal, Heng Ji, Yang Liu, and Dilek Hakkani-Tur. 2022b. Enhancing knowledge selection for grounded dialogues via document semantic graphs. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2810–2823, Seattle, United States. Association for Computational Linguistics.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5081–5087. International Joint Conferences on Artificial Intelligence Organization.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Sewon Min, Jordan Boyd-Graber, Chris Alberti, Danqi Chen, Eunsol Choi, Michael Collins, Kelvin Guu, Hannaneh Hajishirzi, Kenton Lee, Jennimaria Palomaki, et al. 2021. Neurips 2020 efficientqa competition: Systems, analyses and lessons learned. In *NeurIPS 2020 Competition and Demonstration Track*, pages 86–111. PMLR.

Mayank Mishra, Dhiraj Madan, Gaurav Pandey, and Danish Contractor. 2022. Variational learning for unsupervised knowledge grounded dialogs. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4303–4309. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Ashwin Paranjape, Omar Khattab, Christopher Potts, Matei Zaharia, and Christopher D Manning. 2022. Hindsight: Posterior-guided training of retrievers for improved open-ended generation. In *International Conference on Learning Representations*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Neural Information Processing Systems*.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Dmytro Okhonko, Samuel Broscheit, Gautier Izacard, Patrick Lewis, Barlas Ouguz, Edouard Grave, Wen tau Yih, and Sebastian Riedel. 2021. The web is your oyster - knowledge-intensive nlp against a very large web corpus. *ArXiv*, abs/2112.09924.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Peng Qi, Jing Huang, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2021. Conversational ai systems for social good: Opportunities and challenges. *ArXiv*, abs/2105.06457.

Jin Qian, Bowei Zou, Mengxing Dong, Xiao Li, AiTi Aw, and Yu Hong. 2022. Capturing conversational interaction for question answering via global history reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2071–2078, Seattle, United States. Association for Computational Linguistics.

Chen Qu, Liu Yang, Minghui Qiu, W. Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019. Bert with history answer embedding for conversational question answering. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

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*, CHIIR '17, page 117–126, New York, NY, USA. Association for Computing Machinery.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Ori Ram, Gal Shachaf, Omer Levy, Jonathan Berant, and Amir Globerson. 2022. Learning to retrieve passages without supervision. In *NAACL*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

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

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. MobileBERT: a compact task-agnostic BERT for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170, Online. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *ArXiv*, abs/1506.05869.

Ellen M. Voorhees and Dawn M. Tice. 2000. The TREC-8 question answering track. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

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*, pages 38–45, Online. Association for Computational Linguistics.

Donghan Yu, Chenguang Zhu, Yuwei Fang, Wenhao Yu, Shuohang Wang, Yichong Xu, Xiang Ren, Yiming Yang, and Michael Zeng. 2022a. KG-FiD: Infusing knowledge graph in fusion-in-decoder for open-domain question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4961–4974, Dublin, Ireland. Association for Computational Linguistics.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022b. A survey of knowledge-enhanced text generation. *ACM Comput. Surv.* Just Accepted.

Jing Zhao, Junwei Bao, Yifan Wang, Yongwei Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2021. RoR: Read-over-read for long document machine reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1862–1872, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *Computational Linguistics*, 46(1):53–93.

Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2018. Sdnet: Contextualized attention-based deep network for conversational question answering. *arXiv preprint arXiv:1812.03593*.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *ArXiv*, abs/2101.00774.

A Human Evaluation

We used Amazon's Mechanical Turk (AMT) to perform a human evaluation of various models. More specifically, we first created 150 evaluation snippets, each consisting of a dialog context and a set of responses produced by three model variants: LongT5, LongT5-with-Topic, and our full model. Figure 3 shows an example snippet. For each snippet, we asked three different AMT workers to rank the models' outputs based on three criteria: (1) *Fluency*, (2) *Relevance*, and (3) *Interestingness*. Note that for each snippet, the presentation order of the generated responses was randomized. For example, for some AMT tasks, the first response may come from LongT5, while for some other tasks, the first response may come from our full model. Figure 4 shows the instructions we showed at the beginning of each AMT task. Even though our instructions do not explicitly explain how the collected data would be used, we believe it was clear to the AMT workers that we would use the data for researching dialog systems.

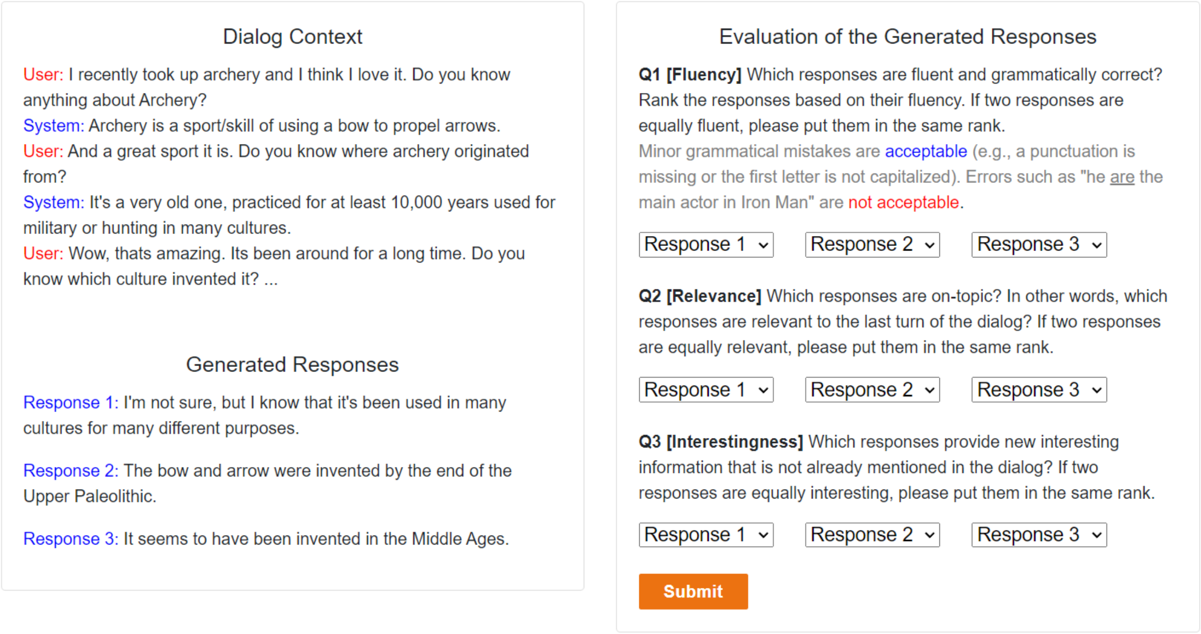
To decide the appropriate cost of each AMT task, we conducted a preliminary study, asking three NLP researchers to do about 20 tasks each. We then computed the average time it would take for each task. After that, we set the value of each AMT task so that someone who worked on our study for about an hour would make at least the required minimum wage.

B Datasets

We use two public datasets in this work: *Wizard of Wikipedia (WoW)* (Dinan et al., 2019) and *MultiDoc2Dial* (Feng et al., 2021). For WoW, we use the version provided by the KILT benchmark (Petroni et al., 2021). KILT is released under the MIT License, and MultiDoc2Dial is released under the Apache-2.0 license. Our use of the datasets is consistent with their licenses.

WoW and MultiDoc2Dial are available at `github.com/facebookresearch/KILT` and `github.com/IBM/multidoc2dial` (respectively). When we first downloaded the datasets, we randomly sampled about 10∼20 examples for each dataset and checked whether the examples contained any offensive content. Overall, we did not see any example that had offensive content.

The two datasets are in English. WoW consists of open-domain conversations, which collectively cover a wide range of topics, ranging from Ameri-



Dialog Context

User: I recently took up archery and I think I love it. Do you know anything about Archery?
System: Archery is a sport/skill of using a bow to propel arrows.
User: And a great sport it is. Do you know where archery originated from?
System: It's a very old one, practiced for at least 10,000 years used for military or hunting in many cultures.
User: Wow, thats amazing. Its been around for a long time. Do you know which culture invented it? ...

Generated Responses

Response 1: I'm not sure, but I know that it's been used in many cultures for many different purposes.

Response 2: The bow and arrow were invented by the end of the Upper Paleolithic.

Response 3: It seems to have been invented in the Middle Ages.

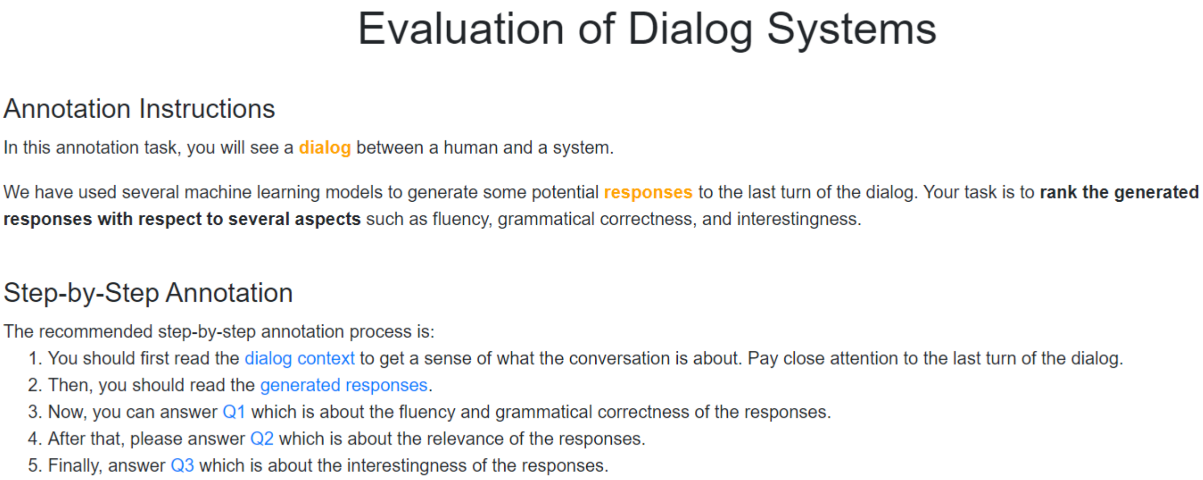
Evaluation of the Generated Responses

Q1 [Fluency] Which responses are fluent and grammatically correct? Rank the responses based on their fluency. If two responses are equally fluent, please put them in the same rank.
Minor grammatical mistakes are acceptable (e.g., a punctuation is missing or the first letter is not capitalized). Errors such as "he are the main actor in Iron Man" are not acceptable.

Q2 [Relevance] Which responses are on-topic? In other words, which responses are relevant to the last turn of the dialog? If two responses are equally relevant, please put them in the same rank.

Q3 [Interestingness] Which responses provide new interesting information that is not already mentioned in the dialog? If two responses are equally interesting, please put them in the same rank.

Figure 3: An example AMT task.



Evaluation of Dialog Systems

Annotation Instructions

In this annotation task, you will see a **dialog** between a human and a system.

We have used several machine learning models to generate some potential **responses** to the last turn of the dialog. Your task is to **rank the generated responses with respect to several aspects** such as fluency, grammatical correctness, and interestingness.

Step-by-Step Annotation

The recommended step-by-step annotation process is:

1. You should first read the dialog context to get a sense of what the conversation is about. Pay close attention to the last turn of the dialog.
2. Then, you should read the generated responses.
3. Now, you can answer Q1 which is about the fluency and grammatical correctness of the responses.
4. After that, please answer Q2 which is about the relevance of the responses.
5. Finally, answer Q3 which is about the interestingness of the responses.

Figure 4: The instructions we provided for each AMT task.

can football to Rock and Roll music. On the other hand, MultiDoc2Dial focuses on four specific domains: Social Security Administration, Veteran Affairs, Student-Aid, and DMV.

If we consider a pair of a dialog context and its corresponding ground-truth response as one example, there are 63,734/3,054/2,944 examples in the train/dev/test splits of WoW (respectively). For MultiDoc2Dial, we refer the readers to the original paper (Feng et al., 2021) for more detailed statistics of the original dataset.

C Reproducibility Information

In this section, we present the reproducibility information of our paper.

Implementation Dependencies Libraries Pytorch (Paszke et al., 2019), Transformers 4.20.1 (Wolf et al., 2020), SentenceTransformers 2.2.0 (Reimers and Gurevych, 2019), faiss-gpu 1.7.2 (Johnson et al., 2021).

Computing Infrastructure We conducted our experiments using Amazon's EC2 virtual machines. Overall, our work can be reproduced using a single p3.8xlarge instance. Information about the cost of using Amazon's EC2 P3 instances can be found at https://aws.amazon.com/ec2/instance-types/p3/.

Number of Model Parameters Our full model consists of three components: a DPR model, a RR model, and a seq2seq generation model. We initialize the DPR model using a Sentence-Transformer model named `all-mpnet-base-v2` (Reimers and Gurevych, 2019), which has about 110M parameters. We initialize the RR model using `distilroberta-base` (Liu et al., 2019), which has about 82M parameters. Finally, we initialize the generation model using `google/long-t5-tglobal-large` (Guo et al., 2022), which has about 750M parameters.

Hyperparameters The effective batch size is 32. The number of training epochs is 10. The base learning rate is 5e-5. K_1 is set to be 25, and K_2 is set to be 5. We use the Adam optimizer and set gradient clipping to 1.0.

Expected Validation Performance For each model variant, we report the test performance of the checkpoint with the best validation F1 score in the main paper (refer to Table 1). On the dev set of WoW, our full model achieves a unigram F1 score of 21.13% and a ROUGE-L score of 18.68%. On the dev set of MultiDoc2Dial, our full model achieves a unigram F1 score of 41.90% and a ROUGE-L score of 37.64%.

# Retrieved Docs	F1	ROUGE-L	Recall@1
$K_1 = 25, K_2 = 5$	38.76	34.62	39.70
$K_1 = 25, K_2 = 10$	39.84	35.81	44.44
$K_1 = 25, K_2 = 15$	39.71	35.58	42.28
$K_1 = 50, K_2 = 5$	39.10	34.92	40.28
$K_1 = 50, K_2 = 10$	39.63	35.44	42.82
$K_1 = 50, K_2 = 15$	39.63	35.56	43.51

Table 7: Results on the dev set of MultiDoc2Dial when we vary K_1 and K_2, the numbers of documents retrieved by DPR and the reranker (respectively). In order to save computational time, we use the base version of LongT5 in this ablation study.

D Additional Analysis

Table 7 shows the results on the dev set of MultiDoc2Dial when we vary K_1 and K_2. In general, increasing the number of retrieved documents typically has marginal positive effects on the evaluation metrics. However, note that increasing the values of K_1 and K_2 also increases the computational complexity and running time of `EKo-DoC`. Therefore, we set K_1 to 25 and K_2 to 5. We believe these values achieve a good balance between the model complexity and the quality of the generated responses.

Table 8 shows some additional examples from WoW that illustrate how incorporating external knowledge can improve the quality of generated responses.

Examples
Topic Document: The first paragraph of the Wikipedia page talking about Ferrari. **Dialog Context**: A: I have 3 Ferrari cars in my garage and I hardly know nothing about them. I read some self help books, then I was rick all of the sudden. B: Those are some nice cars they are manufacturer in Maranello. A: Yeah that's pretty cool. Do you know how old the company is? **Candidate Responses:** LongT5: I don't know but they are a luxury car manufacturer. `EKo-DoC`: I know it was founded in 1939.
Topic Document: The first paragraph of the Wikipedia page talking about Overeating. **Dialog Context**: A: Overeating is a serious problem in the USA. **Candidate Responses:** LongT5: I know, it's a problem that affects a lot of people. `EKo-DoC`: I think it is a serious problem in the US. It is the excess food consumed in relation to the energy that an organism expends.
Topic Document: The first paragraph of the Wikipedia page talking about Red. **Dialog Context**: B: Red is my favorite color. It is the color of a lot of my favorite things, like strawberries? A: Awesome! Did you also know that Mars is red because of iron oxide? ... B: That makes sense. I always liked studying the planets growing up in school. **Candidate Responses:** LongT5-with-Topic: I also like how Mars is the only planet that has a moon. `EKo-DoC`: Yeah, I love learning about the planets. I'm sure you've heard of the planet Mercury.

Table 8: Additional examples showing how external knowledge improves the quality of generated responses. Red is used to indicate responses that are factually *incorrect*. Blue is used to indicate responses that are factually *correct*.