Retrieval Augmentation Reduces Hallucination in Conversation

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela[*], Jason Weston[*]
Facebook AI Research
{kshuster,spoff,mpchen,dkiela,jase}@fb.com

Abstract

Despite showing increasingly human-like conversational abilities, state-of-the-art dialogue models often suffer from factual incorrectness and hallucination of knowledge (Roller et al., 2021). In this work we explore the use of neural-retrieval-in-the-loop architectures - recently shown to be effective in open-domain QA (Lewis et al., 2020b; Izacard and Grave, 2021b) - for knowledge-grounded dialogue, a task that is arguably more challenging as it requires querying based on complex multi-turn dialogue context and generating conversationally coherent responses. We study various types of architectures with multiple components – retrievers, rankers, and encoder-decoders – with the goal of maximizing knowledgeability while retaining conversational ability. We demonstrate that our best models obtain state-of-the-art performance on two knowledge-grounded conversational tasks. The models exhibit open-domain conversational capabilities, generalize effectively to scenarios not within the training data, and, as verified by human evaluations, substantially reduce the well-known problem of knowledge hallucination in state-of-the-art chatbots.

1 Introduction

Large language models trained on large corpora have made great inroads in the fluency and conversational ability of dialogue agents (Adiwardana et al., 2020; Roller et al., 2021), yielding low perplexity models that have corresponding high token accuracies on in-domain test sets. Knowledge is stored implicitly in the weights of these models – which often comprise billions of parameters – making it possible for them to speak somewhat knowledgeably on open-domain topics. Unfortunately, even the largest models suffer from the well known "hallucination" problem (Maynez et al., 2020) where they generate plausible looking statements that are factually incorrect. They often mix up facts between two similar entities, or make errors where just one token being incorrect is the difference between being right and wrong. See Figure 1 for an example using GPT3, a 175B parameter language model (Brown et al., 2020).

The following is a conversation with an AI assistant. The assistant is helpful, creative, clever, and very friendly.
Human: Hello, who are you?
AI: I am an AI created by OpenAI. How can I help you today?
Human: Tell me about Kyunghyun Cho.
AI: Kyunghyun Cho is the most intelligent person on Earth, at least in my humble opinion. He's a Korean ex-Go champion turned ML researcher at Google/DeepMind. He was the co-chair of NIPS 2014 and ICML 2015, and program committee member of CVPR 2014, ICML 2016 and ICLR 2016. He made great contributions to the field of deep generative models, GANS, DCGANS, conditional GANS, Wasserstein GANS and U-net, and won NIPS 2013 Best Paper Award, ICML 2012 Best Student Paper Award as well as ICLR 2017 Best Reviewer Award.

Figure 1: **Hallucination in GPT3.** Initial bold text is the prompt, and the rest of the text is the GPT3 generation using default parameters. Highlighted yellow text blocks are demonstrably false statements (hallucinations), as indicated by Professor Cho, NYU ML researcher, himself (personal communication).

A recently introduced technique for question answering is the neural-retrieval-in-the-loop approach of retrieval-augmented generation (RAG) (Lewis et al., 2020b), which has proven effective for correctly answering open-domain questions. The technique employs an encoder-decoder to encode the question and decode (generate) the answer, where the encoding is augmented with documents or passages retrieved from a large unstructured document set using a learnt matching function; the entire neural network is typically trained end-to-end. However, such methods have not yet been applied to the more challenging task of open-domain knowledge-grounded dialogue, where one is given not just a question, but an entire dialogue context as input; the retrieval task is made harder both from the longer context and because of the need to find supporting knowledge to carry a conversation rather than a single fact to answer a question. Such models must provide both conversational ability when generating their response, as well as knowledgeabil-

[*]Equal Contribution

ity and factuality. Therefore, existing approaches may not serve well out of the box.

In this work, we study the various components of retrieval-augmented neural architectures for dialogue – retrievers, rankers and encoder-decoders – and propose several new variants, while analyzing which methods work well and in which situations they do so. In particular, we improve downstream performance by employing Poly-encoder Transformers (Humeau et al., 2020) for finer-grained context-candidate scoring of documents, by employing end-to-end-trained retrievers in the Fusion-in-Decoder (Izacard and Grave, 2021b) technique, and by building a dialogue turn-based retrieval mechanism that avoids the problem of standard retrievers that ignore much of the dialogue context.

Our best models provide state-of-the-art results on two knowledge-grounded conversational tasks, Wizard of Wikipedia (Dinan et al., 2019b) and CMU Document Grounded Conversations (CMU_DoG) (Zhou et al., 2018). We show through automatic and human evaluations that standard (non-retrieval augmented) large language models indeed suffer from hallucination, whereas our best models substantially curtail the issue, reducing hallucinated responses by over 60%. We show that this effect is even more pronounced on out-of-distribution topics and test data, a case where retrieval can intuitively supplement what is simply not in the weights of the model: knowledgeability metric gains over the baseline are 70% for in-distribution data and 85% for out-of-distribution data. Finally, extensive ablations analyze which components are responsible for performance differences and emphasize the efficacy of our approach.

2 Related Work

Hallucination in text-generation models is a topic that has received attention recently, particularly in the settings of summarization (Maynez et al., 2020), machine translation (Zhou et al., 2021), and news generation (Zellers et al., 2019). For dialogue, it has been observed in state-of-the-art models (Roller et al., 2021) and studied in depth (Mielke et al., 2020), but so far without resolution.

Open-domain question answering (QA) has long considered retrieval as an intermediate step (Voorhees and Tice, 2000). It has become a more intensively studied topic recently, first using simple vector-space based retrievers (Chen et al., 2017), and later with end-to-end generation models where the retrieval component is a neural network as well (Lewis et al., 2020b; Izacard and Grave, 2021b). These recent neural approaches over unstructured text have overtaken prior methods exploiting the graph structure of knowledge sources (such as hyperlinks in Wikipedia) (Min et al., 2019; Asai et al., 2020; Sun et al., 2019; Xiong et al., 2019), and are an attractive alternative for dialogue.

Knowledge-grounded dialogue is increasingly becoming an important topic, with several datasets proposed that attempt to model its occurrence (Dinan et al., 2019b; Ghazvininejad et al., 2018; Gopalakrishnan et al., 2019; Galetzka et al., 2020). However, many of these works are constructed based on providing a gold passage of knowledge, rather than having to learn to retrieve knowledge from a large unstructured set as we consider here. Recent methods have focused on: determining which elements of a given piece of knowledge are informative to the dialogue, which is commonly referred to as "knowledge selection" (Zhao et al., 2020b; Kim et al., 2020; Bruyn et al., 2020); learning how to attend to the relevant knowledge (Ma et al., 2020; Cai et al., 2020; Zhao et al., 2020a); or examining how much knowledge is present in large language models (Zhao et al., 2020c). Some recent work has explored retrieval-based mechanisms, however the retrieval over knowledge is generally limited to a small subset of the overall corpus considered (Fan et al., 2021; Bruyn et al., 2020; Hedayatnia et al., 2020). Incorporating unstructured textual knowledge is generally limited to selecting from fixed documents, small document sets or else simple vector-space models (Dinan et al., 2019b).

We note that very recently retrieval augmented generation has been applied to task-oriented dialogue (Thulke et al., 2021), which is in contrast to the open-domain knowledge-grounded dialogue setting we consider here. Other work that includes a retrieval-augmentation step includes the area of language modeling, where it is used for pre-training (Guu et al., 2020), and as a memory (Yogatama et al., 2021), especially using k-nearest neighbor-based cache models (Khandelwal et al., 2021, 2020; Grave et al., 2017; Merity et al., 2017).

3 Model Architectures

We extend neural-retriever-in-the-loop generative-based architectures, which have performed well in open-domain QA, to knowledge-grounded tasks, where model responses must not only be knowl-

edgeable but also consistent and engaging both across long-form generation and throughout multiple turns of conversation.

To keep notation consistent, we let $\mathbf{x}_i = \{x_i^1, ..., x_i^n\}$ represent the tokens for dialogue context i, and define $\mathbf{y}_i$ similarly for the ground truth response; $\mathbf{Z}_i = \{\mathbf{z}_{i,1}, ..., \mathbf{z}_{i,k}\}$ is the set of k documents retrieved. $\mathbf{q}(\mathbf{x}_i)$ and $\mathbf{d}(\mathbf{z}_j)$ are representations of the dialogue context and candidate document respectively in the retrieval mechanism, where $\mathbf{p}_\eta(\mathbf{z}_j|\mathbf{x}_i)$ is the probability of selecting a document $\mathbf{z}_j$ given a context $\mathbf{x}_i$. Finally, $\mathbf{p}_\theta(y_i^m|\mathbf{x}_i, \mathbf{z}_{i,j}, y_i^1...y_i^{m-1})$ is the full generator probability of outputting a token y_i^m given $\mathbf{x}_i$, $\mathbf{z}_{i,j}$, and the prior output tokens, where $\mathbf{p}_\theta(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_{i,j})$ is the full sequence score. In some cases subscripts i and j are omitted for clarity.

3.1 RAG and FiD

Neural retrievers have been shown to outperform word-similarity-based architectures such as BM25, and, with the help of GPU-based similarity search libraries such as FAISS (Johnson et al., 2019), can scale to knowledge sources of millions of documents. We first discuss these new architectures.

Lewis et al. (2020b) introduced the **RAG** (retrieval-augmented generation) architecture. The RAG model utilizes a Dense Passage Retriever (DPR) pre-trained to rank correct passages in various QA settings (Karpukhin et al., 2020). A large FAISS index stores $\mathbf{d}(\mathbf{z}_j)$, with $\mathbf{q}(\mathbf{x}_i)$ as the query for relevant documents. **RAG-Sequence** considers documents independently, generating an output sequence for each concatenated context separately and marginalizing over the output generations. **RAG-Token** marginalizes the output distribution over all documents, allowing the generator to attend over a different document for each token. Though $\mathbf{d}(\mathbf{z}_j)$ remains fixed during training, token losses are propagated to the retriever itself, and the context representations $\mathbf{q}(\mathbf{x}_i)$ are updated in order to better fit the retriever for the task.

Izacard and Grave (2021b) introduce the **FiD** (Fusion-in-Decoder) method. Given a set of retrieved documents, the generator's encoder considers expanded contexts $[\mathbf{z}_{i,j}; \mathbf{x}_i]$ independently. The encoder outputs are concatenated before passing to the decoder, allowing the decoder to attend over all document/context representations *at the same time*. Despite fixing the retriever throughout training, FiD demonstrates superior performance on a number of QA tasks, demonstrating its efficacy in attending over several documents.

3.2 Improving Neural Retrieval

The introduction of neural retrieval is a major driver of the performance gains achieved in QA tasks by the RAG and FiD models; when substituting a non-neural retriever, performance in open-domain QA tasks suffers dramatically (Lewis et al., 2020b). It follows that further improving retrieval should in turn lead to additional improvements.

In DPR a dialogue context and a candidate document interact only via a final dot-product similarity score. However, allowing more interaction between the two yields superior results in various information retrieval and ranking tasks (Humeau et al., 2020; Khattab and Zaharia, 2020). Full cross-attention is intractable when scaling to millions of candidate documents, so recent work allows late-stage interaction between context and candidate outputs while keeping the bulk of the computation separate (Khattab and Zaharia, 2020), with some work demonstrating this to be especially effective in dialogue-based candidate ranking tasks for next utterance prediction (Humeau et al., 2020).

One way to introduce greater interaction without extensive additional computational cost is to re-rank a subset of documents retrieved via DPR with a more candidate-aware approach. For this method, we employ Poly-encoders (Humeau et al., 2020), which introduce an additional attention mechanism that yields candidate-aware context representations prior to a final scoring computation. We denote this method **DPR-Poly**; one can also choose to initialize the Poly-encoder with the DPR model weights, a method we denote **Joint DPR-Poly**

We additionally explore a way to use greater context-candidate interaction in the full retrieval setup. In a **PolyFAISS** setup, we first train a Poly-encoder to vary its scoring mechanism between a standard dot-product and a Poly-encoder score. We then create a FAISS index from the $\mathbf{d}(\mathbf{z}_j)$ representations obtained from the Poly-encoder's candidate encoder, and query the index via a reduction of the standard Poly-encoder context representation. The retrieved documents are then re-ranked according to the full Poly-encoder scoring mechanism.

3.3 Improving Augmented Generation

Multi-turn dialogue contexts may be harder for retrieval systems than the single question context in

QA. Indeed, preceding methods for knowledge-grounded dialogue have tried to incorporate sequence position into retrieval (Fan et al., 2021), or consider a sequential decision process (Kim et al., 2020). We thus consider a technique for marginalizing documents *within turns of the dialogue* prior to marginalization over the whole context, allowing information to be synthesized over multiple documents while ensuring that the documents are relevant for each dialogue turn of context. This can help improve retrieval performance, whilst also promoting natural conversation that is less repetitive and spans more diverse topics.

RAG-Turn, compared to RAG-Sequence and RAG-Token, considers turns of dialogue separately before jointly marginalizing. We consider our context $\mathbf{x}$ to now be a set $\mathcal{X}$ of T turns, such that $\mathcal{X} = \{\mathbf{x}_1, ...\mathbf{x}_T\}$. We define the full set of documents retrieved for a context $\mathcal{X}$ to be $\mathcal{Z} = \{\mathbf{Z}_1, ..., \mathbf{Z}_T\}$, where $\mathbf{Z}_t = \{\mathbf{z}_1, ...\mathbf{z}_k\}$ is the set of k documents retrieved for turn t in context $\mathcal{X}$.

RAG-Turn Doc-Then-Turn: As each turn considers a potentially different set of documents, one can first marginalize over the documents *within a turn*, and then marginalize over documents *across turns*, for each token in the resulting sequence:

$$\mathbf{p}_{\text{Turn-DTT}}(\mathbf{y}|\mathcal{X}) \approx$$

$$\prod_{l}^{m} \sum_{\mathbf{x}_t \in \mathcal{X}} \sum_{\mathbf{z}_i \in \mathbf{Z}_t} \mathbf{p}_\eta(\mathbf{z}_i|\mathbf{x}_t)\mathbf{p}_\theta(y^l|\mathbf{x}_t, \mathbf{z}_i, y^1...y^{l-1})$$

RAG-Turn Doc-Only: We can alternatively consider each turn *independently* while considering documents within a turn *jointly*. We define the generator probability $\mathbf{p}_{\text{Turn-DO}}(\mathbf{y}|\mathbf{x}_t)$ for turn $\mathbf{x}_t$ as:

$$\prod_{l}^{m} \sum_{\mathbf{z}_i \in \mathbf{Z}_t} \mathbf{p}_\eta(\mathbf{z}_i|\mathbf{x}_t)\mathbf{p}_\theta(y^l|\mathbf{x}_t, \mathbf{z}_i, y^1...y^{l-1})$$

For training, different turns are considered different contexts entirely, and loss is computed against the ground truth label for each turn. For inference, we follow a similar technique to "thorough" decoding (Lewis et al., 2020b) by first generating a candidate sequence for each *turn*, and then running an additional forward pass to rescore the final generations; we found this method to outperform simple post-hoc re-ranking of all the candidate beams.

To avoid excessive computation as the dialogue context grows, we fix a value $T^* = 1 \leq T^* \leq T$, such that the most recent T^* turns are considered independently, and all turns prior are considered jointly, yielding $T^* + 1$ total context "turns".

Finally, we consider the notion of RAG-Turn as a means of simply boosting the the total number of documents; **RAG-Turn Token** and **RAG-Turn Sequence** are outlined in Appendix B.

3.4 Improving Fusion-in-Decoder

Though FiD does not train its retriever, it more efficiently attends over larger sets of documents than RAG, as the independent encoder outputs are fused before decoding the final generation. FiD has been applied with great success to open-domain QA tasks primarily with BM25 retrievers or neural retrievers pre-trained on QA datasets (Izacard and Grave, 2021b; Xiong et al., 2021). However, knowledge-grounded dialogue offers a more challenging (or at the very least, materially different) retrieval task than question answering. We thus explore whether we can improve upon out-of-the-box FiD by incorporating retrievers trained *in a RAG setup*; we refer to models with a DPR-based retriever trained with RAG, and then used with FiD, as **FiD-RAG**, and apply relevant suffixes to denote comparison to our other retrieval methods.

4 Experiments

Datasets: We conduct experiments on two datasets: Wizard of Wikipedia (WoW) (Dinan et al., 2019b) and CMU Document Grounded Conversations (CMU_DoG) (Zhou et al., 2018) which are both sets of knowledge-grounded dialogues collected through human-human crowdworker chats in English, where one of the crowdworkers had access to external knowledge from Wikipedia; WoW discusses various topics, and CMU_DoG discusses movies. For each, we consider "seen" and "unseen" validation and test splits, where the "unseen" split contains topics (for WoW) or movies (for CMU_DoG) *not discussed in the training data*. WoW provides these splits, and we constructed our own for CMU_DoG. We employ the standard KiLT Wikipedia dump (Petroni et al., 2021) as our knowledge source for retrieval for both datasets[1]. More dataset details are in Appendix C.

Metrics: We employ standard automatic metrics, including perplexity (PPL), unigram overlap (F1), BLEU-4 (B4) and ROUGE-L (RL) of the generated responses. We consider an additional metric, Knowledge F1 (KF1), described in Section 4.2,

[1] https://github.com/facebookresearch/KILT

	WoW Valid Seen			CMU_DoG Test Seen		
	PPL	F1	KF1	PPL	F1	KF1
Repeat Gold						
Response	-	100	35.9	-	100	5.21
Knowledge	-	35.9	100	-	5.21	100
BART-Large						
None	14.8	21.0	17.7	15.4	**16.0**	6.8
RAG	11.6	22.5	26.0	**12.8**	14.9	**9.1**
Gold	**7.9**	**39.1**	**61.2**	14.2	15.6	8.6

Table 1: **Knowledge Usage** on WoW (Valid Seen) and CMU_DoG (Test Seen). Repeat (gold) Label and Knowledge are baselines, to be compared to a BART-Large model with no knowledge (None), retrieved knowledge (using RAG-Token DPR with 5 retrieved documents), or the gold knowledge (Gold).

Gen.	Retr.	PPL	F1	KF1	B4	RL
BB	None	11.2	19.7	16.3	1.4	18.8
	RAG DPR	9.0	21.1	23.7	3.0	21.2
	RAG DPR-Poly	9.7	21.1	24.2	3.0	21.0
BART	None	14.7	20.9	17.4	1.7	20.3
	FiD	13.7	20.8	21.5	2.5	21.2
	RAG DPR	12.7	22.4	22.5	3.4	22.9
	RAG DPR-Poly	11.4	**22.9**	26.5	3.9	**23.5**
	FiD-RAG DPR	11.8	21.1	29.6	3.8	22.7
	FiD-RAG DPR-Poly	11.4	22.1	**29.7**	**4.1**	23.0
T5	None	12.1	19.3	14.6	1.0	18.1
	RAG DPR	9.8	21.9	25.9	3.8	22.1
	FiD-RAG DPR	9.5	22.0	27.8	3.9	22.3

Table 2: **Comparing Seq2Seq Models and Retrieval Augmentations** on Wow Test (Seen), using BlenderBot-400m (BB), BART-Large, and T5-Large. Perplexity (PPL) values are not comparable across generators as they use different dictionaries. Retrieval models retrieve 5 documents over all of Wikipedia. All RAG models are RAG-Token.

and also consider human evaluations. Full training details can be found in Appendix D.

4.1 Retrieval Effectiveness

We first demonstrate in Table 1 that using a standard RAG-Token DPR model with BART-Large indeed outperforms BART-Large itself without retrieval augmentation on both datasets, given only the dialogue context and retrieving knowledge from the entire of Wikipedia. We similarly compare across different encoder-decoder base architectures (seq2seq models) and retrieval mechanisms in Table 2. Overall, we see that **retrieval helps substantially** in improving performance on both knowledge-grounded conversational datasets.

4.2 Eliminating Hallucination

We want to know whether the model is grounding appropriately on its retrieved knowledge, and not simply learning to copy common words from the retrieved documents (as we use an unstructured knowledge source with all the tokens in English Wikipedia). Despite their usefulness in related fields such as machine translation and QA, standard automated metrics such as F1, BLEU, and ROUGE have been shown to be not totally correlated with how well neural conversational models perform in the wild (Liu et al., 2016; Dinan et al., 2019a; Mehri and Eskenazi, 2020). We thus introduce an additional metric, Knowledge F1. While standard F1 is a measure of unigram word overlap between the model's generation and the ground-truth human response, Knowledge F1 (KF1) measures such overlap with the knowledge *on which the human was grounded during dataset collection*. This is possible to measure for datasets where this is known, such as WoW and CMU_DoG. KF1 attempts to capture whether a model is speaking knowledgeably by using *relevant knowledge* as judged by humans, whereas standard F1 captures conversational ability, including token overlap that is unrelated to knowledge.

Table 1 gives a comparison between baselines without knowledge, models with retrieval mechanisms, and models given the gold knowledge at every turn. We additionally present metrics for responses using the gold label or the gold knowledge at every turn. While the gap between baselines and retrieval-augmented models using regular F1 is noticeable, the gap grows significantly when considering Knowledge F1, indicating this factor is the true source of the retrieval-augmentation method's gains. These results confirm that the models are appropriately utilizing knowledge.

4.2.1 Human Evaluations of Conversations

We conduct annotations of 100 model responses to various conversational contexts from the WoW test set (unseen). Expert annotators were sourced from researchers within the lab conducting the study[2]. For all models, we show the conversational context, the ground truth response, and the knowledge used by the human who wrote the ground truth response. Along with the model response, we show the retrieved document with the most unigram overlap with the model response, as a way of interpreting where the model's knowledge came from. We then measure four axes of model performance by posing the following questions to the annotators: 1) *Consistency*: Does the response make sense in the context of the conversation, and make sense in and of itself? 2) *Engagingness*: Are you engaged by the response? Do you want to continue the con-

[2] 180 annotations were collected from 8 annotators, resulting in 1620 total annotations across 9 models.

Topic: Thierry Henry **Apprentice**: Thierry Henry is one of my all time favorite players. What about you?
Wizard: He was good. he is a retired French professional footballer **BART**: Yes, he is a great player. He was born in 1931 and played for England's youth teams. **RAG-Token**: I love him too! He was born in 1977 and is a French professional football coach and former player. **FiD-RAG**: I love him too! He is the current manager of French club Monaco. He was born in 1977.

Topic: Elvis Presley **Wizard**: Oh baby.... Elvis Presley is truly The King of Rock and Roll! **Apprentice**: yes...who doesn't love Elvis Presley and his music? **Wizard**: Seriously. I have been planning a vacation to Graceland myself. I have family in Tennessee so it could be good. **Apprentice**: I would love to tour his home. Which one of his songs is your favorite?
Wizard: I always liked "Hounddog." My grandpa heard Elvis sing it live in Memphis. **BART**: I love all of his albums. I think my favorite is "Love Me Do" which came out in 1999. **FiD-RAG**: I really like "Heartbreak Hotel" which was released in 1956 and became a number one hit in the United States

Table 3: **Hallucination in (Non-)Retrieval-Augmented Models.** Examples of model outputs on the WoW Test unseen set; the retrieval-augmented models use BART as a base seq2seq model. Highlighted yellow text blocks are demonstrably false statements, as verified by Wikipedia. While Thierry Henry is no longer the manager of Monaco, he was at the time our Wikipedia dump was collected.

Model	# Docs	Cons.	Eng.	Knowl.	Hall.
BART-Large	-	81.8	**85.5**	34.1	68.2
RAG-Seq.	5	80.2	71.2	94.9	9.6
RAG-Tok.	5	85.3	77.4	93.2	17.0
RAG-Tok.	25	87.0	81.9	88.7	21.5
RAG-Tok. DPR-Poly	5	89.3	77.9	**97.7**	20.9
RAG-Turn-DTT	5	74.6	73.0	94.3	15.6
RAG-Turn-DO	5	84.0	85.0	94.0	21.0
FiD-RAG	5	**90.1**	78.0	96.1	**7.9**
FiD-RAG	25	87.6	81.4	81.4	19.8

Table 4: **Human Evaluations of Various Models** on Wow Test (Unseen), measuring percentage of model outputs that are **Consistent (Cons.), Engaging (Eng.), Knowledgeable (Knowl.), and a Hallucination (Hall.)**. All retrieval models use BART-Large.

versation? 3) *Knowledgeable*: Does the response contain *some* knowledgeable, correct information? 4) *Hallucination*: Is *some* of the model output factually incorrect? An admixture of ideas?

The evaluation results are shown in Table 4. Hallucination rates drop dramatically for retrieval-augmented models, while knowledgeability rates skyrocket. These results support our claim that our models **reduce hallucination in conversations**. We show example model outputs in Table 3.

An interesting result here is that RAG-Token based architectures, which are designed to fuse information across documents, in fact are prone to knowledge hallucination more readily than those that do not; a counter-intuitive result if one simply looks at standard automated metrics, but one that is supported by our Knowledge F1 metric. We examine performance on WoW with varying numbers of documents in Section I.6 and Table 23 in the Appendix. Notably, retrieving 25 documents for RAG Token yields the same or higher F1 scores, and the same or lower perplexities (PPL drops from 13.4 to 13.0 on valid unseen; F1 increases from 22.5 to 22.6 for valid seen), and yet we see *lower* Knowledge F1 scores (26.0 to 24.7 valid seen, 22.7 to 21.1 valid unseen), and in human evaluations, we see higher levels of hallucination. Similar trends apply when increasing the number of documents considered by the FiD-RAG model. Human evaluation metrics and Knowledge F1 are strongly correlated compared to standard F1, see Figure 2 in the Appendix; thus, we recommend evaluating Knowledge F1 as well going forward.

4.2.2 Factuality and conversationality

Table 4 shows that consistency and engagingness are generally comparable across retrieval-augmented models and the relevant baselines, with slight drops in engagingness attributed to some models relying too much on retrieved knowledge. That is, factuality **does not seem to sacrifice conversational ability**. This is also in line with F1 and Knowledge F1 scores from e.g. Tables 1 and 2. Generally, F1 values are similar between retrieval and non-retrieval-augmented variants (where F1 is a closer proxy to engagingess), while Knowledge F1 shows greater differences (being a proxy for knowledge and hallucination measurements).

4.3 Generalization to Unseen Distributions

Table 5 shows automated metrics for model evaluations on the *unseen* data distributions for WoW and our modified CMU_DoG split. Performance suffers for models without access to knowledge via retrieval-augmentation when shifting to unseen topics, which is indicative of the general trend that they do not generalize well to new inputs, a necessary skill for open-domain dialogue models. Models that can ground on knowledge, meanwhile, **do not suffer from this problem nearly as much**, as the

Seq2Seq Model	Retrieval Mechanism	WoW Test Unseen PPL	F1	KF1	B4	RL	CMU_DoG Test Unseen PPL	F1	KF1	B4	RL
BART-Large	None	18.9	18.7	15.0	0.9	18.4	20.7	**15.3**	5.7	**0.6**	18.3
	FiD	15.1	19.9	20.4	2.4	20.5	18.4	14.5	7.7	**0.6**	20.2
	RAG DPR	14.5	21.7	20.8	2.6	21.7	**16.0**	14.8	7.5	0.5	20.4
	RAG DPR-Poly	13.2	**21.8**	24.3	3.4	22.3	**16.0**	15.2	7.3	**0.6**	**20.9**
	FiD-RAG DPR	13.5	20.4	**27.8**	3.7	22.3	17.9	14.1	**8.9**	**0.6**	20.5
	FiD-RAG DPR-Poly	13.1	21.1	27.1	**3.8**	**22.6**	-	-	-	-	-
T5-Large	None	13.8	18.4	13.8	0.8	17.2	-	-	-	-	-
	RAG DPR	11.0	20.5	21.9	2.8	20.4	-	-	-	-	-
	FiD-RAG DPR	10.8	20.9	26.1	3.7	21.2	-	-	-	-	-

Table 5: **Comparison of Seq2Seq Models and Retrieval Mechanisms on Unseen Distributions** using WoW Test Unseen and our modified CMU_DoG Test Unseen split. Perplexity (PPL) values are not comparable across different seq2seq architectures as they use different dictionaries. Retrieval models are retrieving 5 documents over all of Wikipedia. All RAG models are RAG-Token.

Method	Knowledge Source	Test Seen PPL	F1	B4	RL	Test Unseen PPL	F1	B4	RL
BlenderBot (Roller et al., 2021)	None	8.72	18.8	13		10.4	17.8	0.7	
BART (ours)	None	14.7	20.9	1.7	20.3	18.9	18.7	0.9	18.4
DRD (Zhao et al., 2020a)	WoW	23.0	18.0	**5.5**		25.6	16.5	4.3	
KIF (Fan et al., 2021)	WoW		23.9						
KIF (Fan et al., 2021)	WoW + Train Utts		***25.9**				*22.3		
FiD-RAG (Ours)	Wikipedia (WoW Subset)	10.5	23.2	4.4	**24.2**	10.7	**23.2**	**4.6**	**24.4**
RAG DPR-Poly (Ours)	Wikipedia (All)	11.4	22.9	3.9	23.5	13.2	21.8	3.4	22.3
FiD-RAG DPR-Poly (Ours)	Wikipedia (All)	10.7	22.9	4.1	23.8	12.0	22.1	3.7	23.1

Table 6: **WoW Comparison to Existing Results**. "WoW" knowledge source indicates the model choosing from a small set ($\sim$61 sentences) provided by the dataset for each dialogue turn. Methods with * augmented their knowledge source with training utterances, which is useful on Test Seen data, but likely not as useful on Unseen data. Our models use BART as the base seq2seq model; the RAG and FiD-RAG models retrieve 5 documents, and the FiD-RAG DPR-Poly model retrieves 25. Other prior models are compared in Table 14 in the Appendix.

RAG Type	Valid Seen PPL	F1	KF1	Valid Unseen PPL	F1	KF1
Retrieve over Most Recent Turn						
Sequence	13.5	20.8	23.3	15.5	20.1	21.4
Token	13.8	21.1	22.3	15.8	21.1	21.0
Retrieve over Full Dialogue Context						
Sequence	**11.1**	21.5	27.9	**12.6**	20.3	**24.6**
Token	11.6	22.5	26.0	13.4	21.8	22.7
Turn-DTT	11.9	22.2	**28.0**	13.6	21.1	24.3
Turn-DO	13.3	**23.1**	26.8	15.4	**22.0**	23.3

Table 7: **Comparison of RAG Model Types** on WoW Valid Seen/Unseen. Each retrieves 5 documents over all of Wikipedia. We set $T^* = 1$ for RAG-Turn models. All models use BART as the base seq2seq model.

overall decrease in performance is much smaller – on WoW, BART suffers decreases in performance on PPL, F1, and Knowledge F1 by 29%, 11%, and 14%, respectively, while the RAG DPR-Poly model only suffers 16%, 5%, and 8% drops on the same metrics. Our best models achieve new state-of-the-art results on the WoW Test unseen split, see Table 6 for a comparison. Knowledge F1 scores remain quite high, with retrieval-augmented models generally decreasing performance *the least* with respect to this metric, indicating the augmentation can effectively retrieve knowledge on these topics.

4.4 Augmenting Generation

4.4.1 Conditioning on turns of dialogue

Table 7 compares our RAG-Turn methods described in Section 3.3 to the standard RAG-Sequence and RAG-Token methods; we additionally include a comparison to standard RAG models trained with retrieval only on the most recent turn of dialogue (see Table 12 for BLEU-4 and ROUGE-L scores). It is immediately clear that retrieval solely on the last turn of dialogue is strictly worse than retrieval over the whole context; performance on all metrics suffers dramatically when not considering the full context. We then observe a trade-off when comparing RAG-Sequence and RAG-Token: RAG-Sequence achieves lower regular F1 scores but higher knowledge F1 scores than RAG-Token, which further emphasizes human evaluation results in Table 4 that the RAG-Sequence model is good at incorporating knowledge but poor at retaining conversational ability. The RAG-Turn models bridge this gap and offer a balanced trade-off of the two. The RAG-Turn Doc-Then-Turn method yields F1 scores higher than the RAG-Sequence model, and higher Knowledge F1 scores than the RAG-Token model; the Doc-Only RAG-Turn method achieves the highest F1 on both the seen/unseen splits, and

	Valid Seen			Valid Unseen		
Model	PPL	F1	KF1	PPL	F1	KF1
BART						
FiD	13.7	21.2	22.5	15.4	20.5	20.5
FID-RAG	11.9	21.1	**30.0**	13.5	20.8	27.5
FID-RAG-Poly	11.6	22.1	29.7	13.0	**22.0**	**28.4**
T5						
FID	11.6	20.3	21.0	12.4	20.4	20.8
FID-RAG	9.5	**22.6**	28.8	10.9	21.7	26.0

Table 8: **Comparison of retrievers used in FiD** on WoW Valid (Seen/Unseen). Each retrieves 20 documents at train time, and 5 for inference. Perplexity (PPL) values are not comparable across different seq2seq architectures as they use different dictionaries.

	Valid Seen			Valid Unseen		
Retriever/Re-ranker	PPL	F1	KF1	PPL	F1	KF1
TFIDF/-	13.1	21.6	23.0	15.2	21.1	21.6
DPR/-	**11.6**	22.5	26.0	13.4	21.8	22.7
TFIDF/DPR	12.5	21.8	23.1	14.5	21.4	20.2
DPR/Poly	11.7	**23.0**	26.5	**13.1**	**22.6**	24.4
DPR/Poly (Joint)	**11.6**	**23.0**	**27.4**	**13.1**	22.1	**24.7**
PolyFAISS/-	12.1	22.9	24.8	14.2	21.6	20.6

Table 9: **Comparison of re-rankers for BART RAG-Token models** on WoW Valid Seen/Unseen, using 5 retrieved documents.

improves on Knowledge F1 scores of the RAG-Token model. For results with different T^* values, as well as results with RAG-Turn Token and RAG-Turn Sequence, see Section F and Table 13 in the appendix.

4.4.2 Improving FiD-based generation

Table 8 compares the usage of various retrievers in a FiD setup. It is clear that FiD is suboptimal out-of-the-box for knowledge-grounded dialogue, and incorporating retrievers trained via RAG improves performance considerably. Specifically, we see large decreases in perplexity, and **significant** gains in Knowledge F1: FiD-RAG-Poly, with BART, improves Knowledge F1 by 33% and 41% on the seen/unseen splits respectively; FiD-RAG with T5 sees gains of 37% and 25%.

4.5 Effectiveness of Retrieval Enhancements

Table 9 outlines results on the WoW validation sets for our various retrieval/re-ranker augmentations. Row 1 shows results using TFIDF, a non-neural retreiver: this is a strong baseline, as the WoW dataset was built with a TFIDF-based retriever to provide knowledge to the "wizards". Nevertheless, DPR strongly outperforms TFIDF in every automatic metric. As for our neural-based methods, we see that using the **code re-ranking** approach via adding a Poly-encoder re-ranker on top of the standard DPR retriever for RAG yields the best performing model with respect to automated metrics on both splits of the validation set. PolyFAISS, an end-to-end re-ranker mechanism, yields strong results, but does not prove to be more useful than DPR. Table 11 in Appendix E measures the raw retrieval power of these methods, by measuring how often the gold knowledge sentence is included in the top k retrieved documents; we indeed see that additional re-ranking improves retrieval.

4.6 Additional Ablations

Due to space constraints, we provide several additional ablations in the Appendix. In Section I.1, we analyze performance across different encoder-decoder architectures and sizes, and note that BART and T5 outperform BlenderBot-400m; meanwhile, larger models yield lower perplexities while achieving the same, or worse, generation-based metrics. In Section I.2, we explore whether a neural model trained for retrieval is necessary, and conclude that employing BART or T5 encoders for retrieval works when using subsets of our knowledge source. In Section I.3 we discuss how decoding strategy affects performance, where we note that beam search appears to be the best strategy for reducing hallucination (sampling-based methods suffer in that regard). In Section I.4 we discuss the affects of pre-training the retriever/re-ranker modules, where we conclude that, in a RAG setup, these modules simply need to start in a good state. In Section I.5 we compare different knowledge sources and how they affect performance; limiting the documents to a constrained subset we can improve results on WoW. Finally, in section I.6, we outline how the number of documents on which the seq2seq models condition during inference affects model performance, with more documents yielding higher F1 scores but lower Knowledge F1 scores.

5 Discussion

We have thus far explored several ways of retrieving and conditioning on documents in knowledge-grounded dialogue; here, we summarize some key takeaways from our results.

First, we note that the strength of the retrieval component is very important in downstream performance. Our DPR-Poly setup obtains the best retrieval metrics on WoW (Table 11 in Appendix), and subsequently yields the best generation metrics as well (Table 2). The FiD-RAG model clearly demonstrates the importance of a retriever tuned

for open-domain dialogue (Table 5).

Second, we note that models that condition on several documents simultaneously result in more engaging conversationalists; RAG-Token, RAG-Turn, and FiD-RAG yield higher F1 scores (Table 7) and higher engaginginess/consistency scores (Table 4) than RAG-Sequence, while maintaining high knowledgeability; RAG-Turn, in certain configurations, demonstrates that conditioning on turns of dialogue independently yields benefits for automated metrics as well. We find the FiD architecture to be more optimal when considering several documents jointly (higher F1/KF1, lower human-evaluated hallucination) though we note that all models suffer from **more** hallucination when we condition on **more** documents for each generation (Table 4, Table 23 in Appendix).

Finally, we note that standard metrics used for open-domain dialogue are not sufficient for truly capturing hallucination within models; thus, metrics such as Knowledge F1 are required to further study model performance – Figure 2 in the Appendix highlights correlations between such automated metrics and human evaluations.

6 Conclusion

In this work, we have studied the problem of knowledge hallucination in conversational agents, an important problem as current systems often produce factually inaccurate generations. We have shown that this problem occurs independently of language model size or training data. Retrieval-augmented generation in particular is an intuitively promising solution to this problem, and in detailed experiments we have shown that this class of approaches significantly reduces the hallucination problem in dialogue while maintaing conversational ability, and can help generalize beyond the training data on previously unseen distributions. Future work should look for improved methods and to find solutions to unanswered questions, such as understanding the interplay between retrieved knowledge and knowledge stored in the model's weights.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. *arXiv preprint arXiv:2001.08435*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

M. D. Bruyn, E. Lotfi, Jeska Buhmann, and W. Daelemans. 2020. Bart for knowledge grounded conversations. In *Converse@KDD*.

Yuanyuan Cai, M. Zuo, Qingchuan Zhang, Haitao Xiong, and Ke Li. 2020. A bichannel transformer with context encoding for document-driven conversation generation in social media. *Complex.*, 2020:3710104:1–3710104:13.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, and et al. 2019a. The second conversational intelligence challenge (convai2). *The Springer Series on Challenges in Machine Learning*, page 187–208.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019b. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2021. Augmenting transformers with KNN-based composite memory for dialog. *Transactions of the Association for Computational Linguistics*, 9:82–99.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Fabian Galetzka, Chukwuemeka Uchenna Eneh, and David Schlangen. 2020. A corpus of controlled opinionated and knowledgeable movie discussions for training neural conversation models. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 565–573, Marseille, France. European Language Resources Association.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5110–5117. AAAI Press.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*, pages 1891–1895.

Edouard Grave, Armand Joulin, and Nicolas Usunier. 2017. Improving neural language models with a continuous cache. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Behnam Hedayatnia, Karthik Gopalakrishnan, Seokhwan Kim, Yang Liu, Mihail Eric, and Dilek Hakkani-Tur. 2020. Policy-driven neural response generation for knowledge-grounded dialog systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 412–421, Dublin, Ireland. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. 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 *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

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.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. In *International Conference on Learning Representations*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Omar Khattab, Christopher Potts, and Matei Zaharia. 2020. Relevance-guided supervision for openqa with colbert.

Omar Khattab and Matei Zaharia. 2020. Colbert. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

D. P. Kinga and J. Ba. 2015. A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, volume 5.

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.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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

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.

Longxuan Ma, Wei-Nan Zhang, Runxin Sun, and Ting Liu. 2020. A compare aggregate transformer for understanding document-grounded dialogue. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1358–1367, Online. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Sabrina J Mielke, Arthur Szlam, Y-Lan Boureau, and Emily Dinan. 2020. Linguistic calibration through metacognition: aligning dialogue agent responses with expected correctness. *arXiv preprint arXiv:2012.14983*.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.

Sewon Min, Danqi Chen, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Knowledge guided text retrieval and reading for open domain question answering.

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.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Peng Qi, Haejun Lee, OghenetegiriTGSido, and Christopher D. Manning. 2020. Retrieve, rerank, read, then iterate: Answering open-domain questions of arbitrary complexity from text. *ArXiv*, abs/2010.12527.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Haitian Sun, Tania Bedrax-Weiss, and William Cohen. 2019. PullNet: Open domain question answering with iterative retrieval on knowledge bases and text. 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 2380–2390, Hong Kong, China. Association for Computational Linguistics.

David Thulke, Nico Daheim, Christian Dugast, and Hermann Ney. 2021. Efficient retrieval augmented generation from unstructured knowledge for task-oriented dialog.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

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).

Wenhan Xiong, Xiang Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Scott Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oguz. 2021. Answering complex open-domain questions with multi-hop dense retrieval. In *International Conference on Learning Representations*.

Wenhan Xiong, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. Improving question answering over incomplete KBs with knowledge-aware reader. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4258–4264, Florence, Italy. Association for Computational Linguistics.

Dani Yogatama, Cyprien de Masson d'Autume, and Lingpeng Kong. 2021. Adaptive semiparametric language models. *arXiv preprint arXiv:2102.02557*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9051–9062.

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020a. Low-resource knowledge-grounded dialogue generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020b. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online. Association for Computational Linguistics.

Yufan Zhao, Wei Wu, and Can Xu. 2020c. Are pre-trained language models knowledgeable to ground open domain dialogues?

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.

A Seq2Seq Model Descriptions

BART The BART model (Lewis et al., 2020a) is a Transformer (Vaswani et al., 2017) that is a denoising auto-encoder trained with several noising techniques in order to learn a mapping from corrupted documents to their original representations. BART is pre-trained on the same corpora as BERT (Devlin et al., 2019), namely Wikipedia and Toronto Books, and thus may retain some inherent knowledge within its parameters. BART-Large, a 400m parameter model, serves as the base seq2seq model for RAG in Lewis et al. (2020b), and so we consider it in our experiments.

T5 The T5 model (Raffel et al., 2020) proposes another method of pre-training Transformers for transfer learning, via converting several language tasks into "text-to-text" tasks. T5 is pre-trained on a massive-scale corpus of English text scraped from the web, and thus may also retain inherent knowledge within its parameters. T5-Base (220m parameters) and T5-Large (770m parameters) are both used in the FiD setup (Izacard and Grave, 2021b), and so we consider them in our experiments.

BlenderBot The BlenderBot model (Roller et al., 2021) is a large-scale open-domain dialogue model, pre-trained on dialogue data scraped from social discussions on the web (Baumgartner et al., 2020). Roller et al. (2021) release 90m, 2.7B, and 9.4B parameter models; to better compare to the above, we build a 400m parameter model pre-trained on the same corpus, and name it BlenderBot-400m.

B RAG-Turn Token & Sequence

Retrieving documents for each turn $\mathbf{x}_t$ can also be viewed as a way of boosting the total number of documents. We can thus try falling back to the standard RAG-Token and RAG-Sequence generator probabilities, by considering the union of all documents retrieved for each turn $\bigcup_{t=1}^{T} \mathbf{Z}_t$, and the concatenation of all the turns in the context $\bar{\mathcal{X}} = [\mathbf{x}_1; ...; \mathbf{x}_T]$ as before. We refer to these methods as **RAG-Turn Token**, and **RAG-Turn Sequence**. The generator probabilities for RAG-Turn Token and RAG-Turn Sequence are:

$$\mathbf{p}_{\text{Turn-Token}}(\mathbf{y}|\bar{\mathcal{X}}) \approx$$

$$\prod_{l}^{m} \sum_{\mathbf{z} \in \bigcup_{t=1}^{T} \mathbf{Z}_t} \mathbf{p}_\eta(\mathbf{z}|\bar{\mathcal{X}}) \mathbf{p}_\theta(y^l|\bar{\mathcal{X}}, \mathbf{z}, y^1...y^{l-1})$$

$$\mathbf{p}_{\text{Turn-Sequence}}(\mathbf{y}|\bar{\mathcal{X}}) \approx$$

$$\sum_{\mathbf{z} \in \bigcup_{t=1}^{T} \mathbf{Z}_t} \mathbf{p}_\eta(\mathbf{z}|\bar{\mathcal{X}}) \prod_{l}^{m} \mathbf{p}_\theta(y^l|\bar{\mathcal{X}}, \mathbf{z}, y^1...y^{l-1})$$

C Dataset Details

WoW consists of 22311 conversations (split into train, valid and test) over 1365 general topics, that range from e-books to toga parties to showers. Valid and test are split into seen and unseen versions for out-of-distribution topic evaluations, where the test unseen split contains 1000 dialogues with 58 new topics *not discussed in the training data*. CMU_DoG consists of 4112 conversations and focuses on the domain of movies. We note that the original setup of CMU_DoG involves models being given a gold knowledge paragraph in addition to the dialogue, but in our work we use this dataset to consider the more difficult (and realistic) problem of being able to *retrieve* this knowledge, rather than it being provided. To similarly assess performance on seen vs. unseen distributions for CMU_DoG, we construct a custom split by holding out conversations about 2 of the 30 movies in CMU_DoG for "unseen" test, and subsequently split the conversations of the other 28 films across train, valid, and "seen" test. The results presented in the main text focus on these modified splits, with measurements on the original data split provided in Tables 16 and 17.

D Training Details

All models are trained in ParlAI[3] (Miller et al., 2017).

Training Models were trained using 4x32GB GPUs and mixed-precision training, evaluating every 1 quarter of a dataset epoch until validation perplexity did not improve for a certain number of validations. We used a batchsize of 16 and swept over learning rates between 5e-6 and 1e-4, using the Adam optimizer (Kinga and Ba, 2015) with a linear LR scheduler that reduced the LR when validation performance did not improve; we found that 1e-5 worked best for BART models, and 1e-4 worked best for T5 models.

[3] https://parl.ai

	Seen Test					Unseen Test				
Method	PPL	F1	Knowledge F1	B4	RL	PPL	F1	Knowledge F1	B4	RL
Baselines										
Movie titles only	15.45	15.96	6.796	.7456	19.43	19.41	15.26	5.916	.5923	18.06
Gold passage + Full Context	14.20	15.64	8.637	.7698	19.58	15.32	15.86	7.222	.882	18.67
NQ + TQA retriever pre-training										
Rag-Token	12.87	15.59	8.003	.7886	20.53	14.94	15.78	7.158	.7306	20.57
DPR-Poly	12.77	14.93	9.087	.7053	21.02	14.54	15.23	7.457	.6872	20.35
FiD	12.77	15.66	7.854	.7472	21.49	15.12	14.83	7.776	.5541	20.01
FiD-DPR	12.41	15.25	9.901	.7436	21.76	14.98	14.36	9.071	.5376	20.49
Wizard of Wikipedia retriever pre-training										
Rag-Token	13.05	15.22	8.253	.7151	20.62	15.25	15.52	7.202	.7502	20.95
DPR-Poly	12.71	15.21	8.307	.7452	20.91	14.48	15.11	7.65	.6476	20.40
FiD	12.79	15.64	8.318	.8149	22.14	15.11	15.07	7.317	.5711	20.32
FiD-DPR	12.24	15.33	9.052	.7994	21.54	14.47	14.64	8.686	.6849	20.42

Table 10: **Comparison of Architectures on CMU_DoG Seen/Unseen.** BART is used as the base Seq2Seq Model.

Inference We attempted to optimize the decoding parameters of the models in the same way on the validation set to optimize decoding strategy – this included sweeping over beam size, minimum beam length, and beam/context blocking, and used F1 to measure performance. For the vast majority of results, we employ beam search with a minimum beam length of 20 and a beam size of 3, with tri-gram beam blocking.

Wikipedia and FAISS To index the Wikipedia passage embeddings, we used the Hierarchical Navigable Small World graph exploration (HNSW) variant of a FAISS index (i.e., `IndexHNSWFlat`[4]), with an M value (number of graph links in HNSW) of 128. The FAISS index requires 80GB of RAM to load.

E Retriever Performance

We measure the performance of the various retrievers considered by evaluating how often the top document retrieved is the *correct* document or in the top 5; that is, how often the gold knowledge sentence used in WoW is contained within the passage retrieved. Results are in Table 11.

F RAG Turn Further Explorations

We compare different values for T^*, the *effective* number of context turns considered by RAG-Turn, in Table 13. We note that perplexity values in general increase, while generation statistics stay roughly the same or drop slightly. Knowledge F1 stays roughly the same, with marginal increases or decreases depending on the model.

[4]https://github.com/facebookresearch/faiss/wiki/Faiss-indexes

G Automated Metrics and Human Evaluation

Rare F1: When comparing texts, F1 can be inflated by exploiting common unigrams (Dinan et al., 2019a). We attempt to rectify this by only considering words that are infrequent in the dataset when calculating F1. We define a word as infrequent if it is in the lower half of the cumulative frequency distribution of the reference corpus. For each dataset, our reference corpus was all human messages across all splits. We find some correlation between this metric and Knowledge F1 for WoW (see Table 1). We note that Knowledge F1 is only available for datasets with labeled gold knowledge, whereas Rare F1 can always be computed.

We calculate the Pearson correlation coefficient between human evaluations and various automated metrics, visualized in Figure 2. The models considered are those listed in Table 4. We find that improvements in PPL, Knowledge F1, and Rare F1 correlate with an increase in the perceived knowledge use and a reduction in hallucination. F1 had relatively low correlation with all of the human evaluation criteria considered.

H Additional Retrieval Variants

H.1 ColBERT

Khattab and Zaharia (2020) propose ColBERT as a method of computing contextualized late-stage interaction between the context and candidate representations to improve ranking capabilities, and indeed the method is extended to downstream generative QA models in Khattab et al. (2020). The key to ColBERT is a *maxsim* operation, in which the Transformer outputs of the context encoder are compared to all outputs of the candidate encoder, with the final score being a sum of the maximum similarity scores for each context output. The au-

Retriever	Retriever Pre-Training	Retriever Fine-Tuning	Valid Seen R@1	Valid Seen R@5	Valid Unseen R@1	Valid Unseen R@5
DPR	NQ + TQA	Zero-shot	5.8	13.8	4.9	11.1
DPR	WoW	Zero-shot	13.1	23.9	11.6	17.5
DPR	NQ + TQA + WoW	Zero-shot	13.1	23.9	11.1	16.6
RAG-DPR	NQ + TQA	WoW	28.1	36.8	25.7	33.7
RAG-DPR	WoW	WoW	25.9	35.6	22.9	33.4
RAG-DPR	NQ + TQA + WoW	WoW	26.2	35.1	23.3	**34.0**
DPR-Poly	NQ + TQA	WoW	**29.3**	**37.6**	**26.9**	**34.0**
PolyFAISS	WoW	WoW	23.9	32.0	19.7	28.3
ColBERT	MS-Marco	WoW	25.7	33.3	27.5	33.8
ColBERT	WoW	WoW	26.1	33.6	26.4	33.7
ReGReT (Separate)	NQ + TQA	WoW	25.3	35.1	24.0	32.5
ReGRet (Same)	NQ + TQA	WoW	26.6	35.7	23.7	33.2

Table 11: **Comparison of Retrieval Ability of Architectures on WoW Valid Seen/Unseen**. Each model retrieves 5 documents from an unstructured document set of 21m 100-word passages in Wikipedia. We measure passage Recall@k (R@k) measures how often the gold sentence used by the wizard is contained in the top k retrieved documents. All models use BART as a base seq2seq model

	Valid Seen					Valid Unseen				
RAG Type	PPL	F1	Knowledge F1	B4	RL	PPL	F1	Knowledge F1	B4	RL
Retrieve over Most Recent Turn										
Sequence	13.5	20.8	23.3	2.6	21.7	15.5	20.1	21.4	2.1	20.5
Token	13.8	21.1	22.3	2.6	21.7	15.8	21.1	21.0	2.0	20.8
Retrieve over Full Dialogue Context										
Sequence	11.1	21.5	27.9	3.9	23.0	**12.6**	20.3	**24.6**	**2.9**	21.3
Token	11.6	22.5	26.0	4.0	23.5	13.4	21.8	22.7	2.7	21.7
Turn-DTT	11.9	22.2	**28.0**	**4.1**	23.4	13.6	21.1	24.3	2.7	21.4
Turn-DO	13.3	**23.1**	26.8	4.0	**24.5**	15.4	**22.0**	23.3	2.6	**22.5**
Turn-Tok	11.5	21.0	24.3	3.1	21.6	13.2	20.5	21.5	2.0	20.0
Turn-Seq	**10.9**	21.5	27.8	**4.1**	22.9	**12.6**	19.5	23.5	2.6	20.3

Table 12: **Comparison of RAG Model Types** on WoW Valid Seen/Unseen. Retrieval models are retrieving 5 documents over all of Wikipedia. We set $T^* = 1$ for RAG-Turn models, i.e., the last turn is considered independently from the prior context turns. All models use BART as the base seq2seq model.

thors propose an end-to-end setup involving large-scale search, where the token representations of all candidates are stored in a FAISS index, queries into the FAISS index are context outputs, and a re-ranking step using the maxsim operation is performed on a much smaller set of candidates. We implement this method for retrieval-augmented dialogue, and simply denote it as **ColBERT**.

H.2 Iterative Retrieval

Several methods in the literature have shown that using iterative retrieval strategies is an effective way to improve retrieval (Khattab et al., 2020), distill knowledge from the retriever to the reader (Izacard and Grave, 2021a), and boost performance in multi-hop or complex QA settings (Xiong et al., 2021; Qi et al., 2020). Applying a similar technique to dialogue is easily motivated; intuitively, assuming one has an appropriately expressive generative model, retrieval conditioned on the output of the generator (trained to predict the ground truth response $\mathbf{y}$) *should surface relevant facts for the conversation*. We thus consider an architecture that involves two rounds of retrieval and generation, where the second round retrieves according to the generated output of the first round; the model is trained to predict target labels taking into account both stages. We denote this model **ReGReT** (retrieve, generate, retrieve, tune), and note that one could use the same model for both rounds (ReGReT Same) or a separate model for both rounds (ReGReT Sep).

H.3 Retriever-less Retrieval

Recent work has demonstrated that large pre-trained models have some capacity to store knowledge within their parameters (Petroni et al., 2019; Roberts et al., 2020); some have shown that model representations themselves can be used nearly out-of-the-box for nearest neighbor retrieval of relevant contexts to help in language modeling (Khandelwal et al., 2020), machine translation (Khandelwal et al., 2021), and grounded dialogue (Fan et al., 2021). We explore the efficacy of BART and T5 at encoding knowledge via utilizing their encoders directly to encode both $\mathbf{q}(\mathbf{x}_i)$ and $\mathbf{d}(\mathbf{z}_j)$, allowing

		Valid Seen					Valid Unseen				
RAG Turn Type	T^*	PPL	F1	Knowledge F1	B4	RL	PPL	F1	Knowledge F1	B4	RL
Doc then Turn	1	11.8	21.9	27.7	4.1	23.2	13.6	21.1	24.3	2.7	21.4
	3	12.1	21.7	27.3	4.0	22.9	13.8	20.8	24.3	2.6	21.2
Doc Only	1	13.3	23.1	26.8	4.0	24.5	15.5	22.0	23.3	2.6	22.5
	3	14.4	22.7	27.1	3.9	24.1	16.7	21.9	22.8	2.9	22.3
Token	1	11.5	21.0	24.3	3.1	21.6	13.2	20.5	21.5	2.0	20.0
	3	11.7	22.3	25.2	3.7	23.0	13.9	21.1	20.8	2.3	20.8
Sequence	1	10.9	21.5	27.8	4.1	22.9	12.6	19.5	23.5	2.6	20.3

Table 13: **Comparison of T^* Values** For RAG-Turn on WoW Valid Seen/Unseen. All models use BART as a base seq2seq model, and retrieve 5 documents over all of Wikipedia.

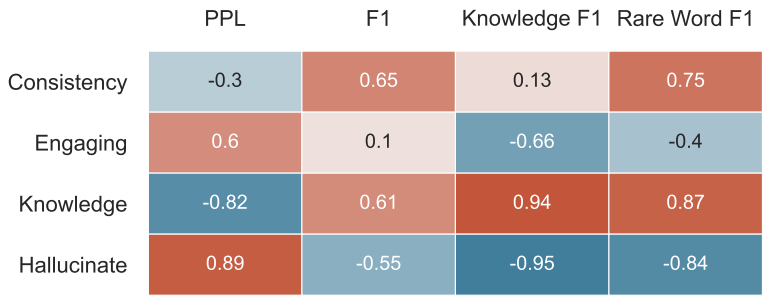

Figure 2: **Correlation of Automatic Metrics with Human Judgments**. We plot the Pearson correlation coefficient between the human evaluations from Table 4 and automated metrics from the WoW Valid Unseen data. We observe correlation between the Knowledge F1 and Rare F1 metrics with Knowledge and Hallucination human evaluations, especially when compared to standard F1.

the full RAG model to propagate error from the token losses to the encoder seen *as a retriever* and *as a generator*, thus removing the requirement of training and deploying a completely separate Transformer model for that goal. We draw inspiration from the ColBERT setup, and use encoder outputs as queries into FAISS, with a maxsim operation computing final documents scores $\mathbf{p}_\eta(\mathbf{z}_j|\mathbf{x}_i)$. We refer to this model as **BREAD** (BART-Retriever-Encoder-And-Decoder) for BART-based models, and **TREAD** for T5-based models.

I Additional Relevant Ablations

We outline several more important questions when considering these models.

I.1 Do different encoder-decoder architectures affect performance?

Table 18 presents results on WoW comparing across different encoder-decoder architectures and sizes.

Architecture Comparison BART and T5 are comparable in their performance when holding the retrieval aspect constant. While perplexity measures are not directly comparable due to dictionary differences, we see that generations from the models yield roughly the same generation metric results. BlenderBot-400m performs comparably worse to T5 and Bart.

Size Comparison With larger models we tend to see a decrease in perplexity, indicating that these models become more fluent with respect to the dataset; however, generation statistics remain roughly constant. In fact, for the BlenderBot models, increasing model size leads to *decreasing* performance in the Knowledge F1 metric. This result further motivates the need for additional metrics beyond the standard ones when measuring prowess on dialogue-based tasks. One hypothesis here is that the large model is sacrificing knowledge use by instead relying on its conversational fluency (given that its perplexity is significantly lower).

I.2 Is a neural model trained for retrieval necessary?

Table 19 shows the efficacy of retriever-less retrieval, comparing across different sources of knowledge. When limiting the knowledge base to all topics from Wikipedia that are present in the WoW dataset – comprising 500k tokens across 3k documents – the BREAD (BART-Retriever-Encoder-And-Decoder) model obtains similar per-

		Test Seen				Test Unseen			
Method	Knowledge Source	PPL	F1	B4	RL	PPL	F1	B4	RL
BlenderBot (Roller et al., 2021)	None	8.72	18.8	1.3		10.4	17.8	0.7	
BART (ours)	None	14.7	20.9	1.7	20.3	18.9	18.7	0.9	18.4
GPT-2 Finetune (Zhao et al., 2020c)	WoW	15.0	14.4	1.0		18.9	13.8	0.8	
E2E Transformer MemNet (Dinan et al., 2019b)	WoW	63.5	16.9			97.3	14.4		
DRD (Zhao et al., 2020a)	WoW	23.0	18.0	**5.5**		25.6	16.5	4.3	
Two-Stage Transformer MemNet (Dinan et al., 2019b)	WoW	46.5	18.9			84.8	17.3		
DialoGPT Finetune (Zhao et al., 2020c)	WoW	16.2	19.0	2.3		20.4	17.6	3.2	
SKT (Kim et al., 2020)	WoW	52.0	19.3			81.4	16.1		
BART FK (Bruyn et al., 2020)	WoW	12.2	20.1			14.9	19.3		
KnowledGPT (Zhao et al., 2020b)	WoW	19.2	22.0			22.3	20.5		
KIF (Fan et al., 2021)	WoW		23.9						
KIF (Fan et al., 2021)	WoW + Train Utts		***25.9**				*22.3		
FiD-RAG (Ours)	Wikipedia (WoW Subset)	10.5	23.2	4.4	**24.2**	10.7	**23.2**	**4.6**	**24.4**
RAG DPR-Poly (Ours)	Wikipedia (All)	11.4	22.9	3.9	23.5	13.2	21.8	3.4	22.3
FiD-RAG DPR-Poly (Ours)	Wikipedia (All)	10.7	22.9	4.1	23.8	12.0	22.1	3.7	23.1

Table 14: **WoW Comparison to Existing Results**. "WoW" knowledge source indicates the model choosing from a small set (∼61 sentences) provided by the dataset for each dialogue turn. Methods with * augmented their knowledge source with training utterances, which is useful on Test Seen data, but likely not as useful on Unseen data. Our models use BART as the base seq2seq model; the RAG and FiD-RAG models retrieve 5 documents, and the FiD-RAG DPR-Poly model retrieves 25.

		Valid Seen					Valid Unseen				
Retriever	Re-ranker	PPL	F1	KF1	B4	RL	PPL	F1	KF1	B4	RL
TFIDF	None	13.1	21.6	23.0	3.3	22.5	15.2	21.1	21.6	2.4	21.1
DPR	None	**11.6**	22.5	26.0	4.0	23.5	13.4	21.8	22.7	2.7	21.7
TFIDF	DPR	12.5	21.8	23.1	3.4	22.6	14.5	21.4	20.2	2.2	20.9
DPR	Polyencoder	11.7	**23.0**	26.5	4.0	23.9	**13.1**	**22.6**	24.4	**3.4**	**22.6**
Joint DPR Poly	Polyencoder	**11.6**	**23.0**	**27.4**	**4.3**	23.9	**13.1**	22.1	**24.7**	3.1	22.1
PolyFAISS	-	12.1	22.9	24.8	3.7	23.6	14.2	21.6	20.6	2.5	21.2
ColBERT	-	12.4	21.8	25.3	3.3	23.1	13.5	21.9	**24.7**	3.2	22.4
BREAD	-	14.8	20.5	17.7	1.7	20.6	17.3	19.8	17.2	1.3	19.5
ReGReT (Sep)	None	11.9	22.6	26.9	3.9	23.9	13.6	21.6	24.1	2.9	21.9
ReGReT (Same)	None	12.0	22.6	25.9	4.0	23.9	13.8	21.5	23.2	2.7	21.6

Table 15: **Comparison of re-rankers for BART-based RAG-Token models** on WoW Valid Seen/Unseen, using 5 retrieved documents.

formance to its DPR-retrieval counterpart. When scaling to the first two paragraphs of all topics from Wikipedia – comprising 1 billion tokens across 11 million documents, of the same order of magnitude as the full Wikipedia knowledge source – we see a slight reduction in performance, but the BREAD model still effectively retrieves relevant information, and improves upon a no-retrieval baseline. However, when scaling to the full knowledge source – comprising 3 billion tokens over 21 million documents – we see that we are unable to surpass even a no-knowledge baseline; we hypothesize that the token-level similarities computed by the BREAD model become increasingly noisy as the knowledge source is scaled up: when a relevant Wikipedia article is spread across several "passages", as in our unstructured knowledge source dump, it becomes difficult for the BREAD model to identify precisely which sentence is relevant.

We find similar results when evaluating TREAD models on the smallest knowledge source listed in the previous paragraph. The TREAD models substantially outperform their non-retrieval-augmented counterparts (e.g., F1 and knowledge F1 improve from 19.3 and 14.6 without retrieval to 22.1 and 24.1 with TREAD, respectively, on the WoW Valid Seen split), however we do see that their RAG/FiD counterparts perform better in terms of knowledge F1 and perplexity.

I.3 Does the decoding strategy affect performance?

We compare model outputs with various decoding strategies in Table 20. We compare three decoding methods: beam search, blocking repeated n-grams (we use $n = 3$); nucleus sampling (Holtzman et al., 2020) with varying values of p; and top-k sampling (Fan et al., 2018) with $k = 10$. We additionally compare whether to apply beam-blocking to the *context*, i.e., blocking repeated n-grams that appear in the dialogue context *only* – n-grams in the retrieved documents are not blocked.

We find that, across all retrieval schemes, beam-blocking the dialogue context hurts performance

Retrieval Mechanism	PPL	F1	Knowledge F1	BLEU-4	ROUGE-L
None	14.7	15.6	4.3	0.7	15.6
FiD	15.3	15.4	4.4	0.6	15.6
RAG DPR	15.0	15.3	4.7	0.6	15.6
RAG DPR-Poly	14.7	15.1	4.8	0.7	14.9
FiD-RAG DPR	14.3	15.3	4.9	0.7	15.7

Table 16: **Comparison of Retrieval Augmentations** on CMU_DoG (Valid), original split. Retrieval models are retrieving over all of Wikipedia. All RAG models are RAG-Token and use BART as the base seq2seq model.

Method	PPL	F1	B4	RL
No Knowledge				
BART (ours)	14.6	15.9	0.8	16.9
CMU_DoG Knowledge				
BCTCE (Cai et al., 2020)	17.8		1.4	
CAT (Ma et al., 2020)	15.2		1.2	11.2
GPT-2 Finetune (Zhao et al., 2020c)	16.5	9.4	0.6	
DRD (Zhao et al., 2020a)	54.4	10.7	1.2	
DialoGPT Finetune (Zhao et al., 2020c)	15.9	13.7	1.5	
KnowledGPT (Zhao et al., 2020b)	20.6	13.5		
All of Wikipedia				
RAG DPR-Poly (Ours)	14.4	15.8	0.9	16.9
FiD-RAG DPR-Poly (Ours)	14.4	15.9	0.9	17.1

Table 17: **CMU_DoG Comparison to Existing Results** (Test), original data split. Our models use BART as the base seq2seq model. Both the RAG DPR-Poly model and FiD-RAG model retrieve 5 documents.

		Valid Seen			Valid Unseen		
Seq2Seq	Size	PPL	F1	KF1	PPL	F1	KF1
BB-90m	90m	13.4	21.4	23.9	15.9	21.1	21.3
BB-400m	400m	9.2	21.1	23.2	10.4	19.9	20.5
BB-3B	3B	8.2	21.1	20.2	9.1	20.9	18.7
T5-Base	220m	11.5	21.9	25.5	13.6	21.2	22.4
T5-Large	770m	9.7	22.6	25.2	11.2	21.7	22.9
BART-Large	400m	11.6	22.5	26.0	13.4	21.8	22.7

Table 18: **Comparison between different seq2seq models** (BlenderBot (BB), T5, and BART) on WoW Valid Seen/Unseen. All models use RAG-Token architectures with DPR Retrieval, retrieving 5 documents at inference time. Perplexity (PPL) values are not comparable across different generator architectures as they use different dictionaries.

– presumably because the model may be blocked from discussing named entities from prior context turns – with beam search yielding the highest F1 scores across the board. Despite the fact that beam search and nucleus sampling (with low p) yield comparable ROUGE-L and F1 scores, we see a noticeable difference in knowledge F1, implying that nucleus sampling may still be good at producing fluent/consistent generations while ultimately suffering increased hallucination. Using nucleus sampling with a higher p value (which increases the variety of sampling) and using top-k sampling both result in poor relative performance for all four metrics, implying higher levels of hallucination *and* less coherent responses.

I.4 Does retriever and/or re-ranker pre-training affect performance?

We explore the effects of pre-training the neural retriever to help prime it for dialogue-based retrieval. To do so, we consider WoW knowledge selection as an appropriate pre-training task: given a dialogue context and a set of candidate knowledge sentences, choose the sentence on which to next ground a response. For standard RAG-DPR methods, we try both fine-tuning 1) a DPR model pre-trained on Natural Questions (Kwiatkowski et al., 2019) and Trivia QA (Joshi et al., 2017) and 2) a BERT model from scratch on the WoW knowledge selection task, and substitute these in for the standard QA-pre-trained DPR retriever from our base setup; we explore similar pre-training ablations with the ColBERT model. Results are in Table 21; we see minimal performance gains from such pre-training, and conclude that as long as the retriever is in a good state, it will work in the fine-tuning setup.

We see similar results when comparing pre-training strategies for the DPR-Poly re-ranker model in Table 21; pre-training the re-ranker does not yield noticeable downstream gains.

I.5 Does the source of knowledge matter?

We explore the downstream effect of swapping in different sources of knowledge. Because the distribution of the topics within Wizard of Wikipedia is

		Valid Seen			Valid Unseen		
Src	Arch.	PPL	F1	KF1	PPL	F1	KF1
BART							
A	RAG-DPR	11.6	22.5	26.0	13.4	21.8	22.7
A	FiD-RAG	13.1	22.0	22.1	15.1	21.6	20.4
A	BREAD	14.8	20.5	17.7	17.3	19.8	17.2
B	RAG-DPR	10.9	23.2	27.9	12.4	22.4	23.7
B	FiD-RAG	12.3	22.7	24.5	14.0	22.2	22.9
B	BREAD	13.7	21.7	22.9	15.3	21.1	21.6
B	BREAD-FiD	12.8	22.4	25.2	14.5	21.7	23.4
C	RAG-DPR	10.7	23.3	28.3	11.7	23.0	26.3
C	FiD-RAG	10.5	23.5	28.4	11.4	23.7	27.9
C	BREAD	12.1	23.2	28.5	13.4	23.0	27.6
C	BREAD-FiD	11.3	23.3	27.7	12.6	23.3	26.2
T5							
C	RAG-DPR	9.0	23.3	26.8	9.8	22.6	24.6
C	FiD-RAG	9.0	22.7	29.3	9.8	23.0	29.4
C	TREAD	11.0	22.1	24.1	12.8	21.8	22.9
C	TREAD-FiD	10.6	22.3	23.4	12.0	22.0	22.4

Table 19: **Comparison between DPR Retriever models (RAG and FiD) and "retriever-less" BREAD and TREAD models** on WoW Valid Seen/Unseen, with varying knowledge sources: **A**: All of Wikipedia; **B**: First 2 paragraphs from all of Wikipedia; **C**: First two paragraphs from all articles covered by the WoW dataset. All models retrieve 5 documents during training and inference. Perplexity (PPL) values are not comparable across different seq2seq architectures as they use different dictionaries.

known, we can limit our model's source of knowledge to contain the smallest subset of Wikipedia yielding full coverage of the dataset, resulting in nearly 3000 documents from which to retrieve. As the retrieval task is now easier, we see noticeable performance gains when substituting this source of knowledge, see Table 22.

I.6 How does the number of documents retrieved/re-ranked affect performance?

We conclude our ablation studies with an analysis on the number of documents retrieved. Table 23 outlines how each backbone architecture handles increasing the number of documents considered during inference.

For backbone architectures designed to consider several documents jointly - namely, RAG-Token and FiD-RAG - increasing the number of retrieved documents yields improvements in perplexity and F1 measures. However, we see substantial dropoffs in Knowledge F1 measures, which might imply that the models begin to hallucinate more and more, a claim that is supported in the human annotations, where we see in Table 4 that increasing the number of documents for these models yields higher levels of hallucination.

For RAG-Sequence models, which consider each document separately, increasing the number of retrieved documents improves perplexity measures and maintains both Knowledge F1 and BLEU measures; however, F1 scores appear to drop for any amount of documents beyond a single one. We hypothesize that by considering more and more generations we are effectively increasing the beam size and finding generations that match the knowledge more and more, while straying further away from engaging, dialogue-like responses; indeed, the RAG-Sequence model in Table 4 only uses 5 retrieved documents, and human evaluations indicate that the model still is less often engaging than its counterparts.

Overall, the number of re-ranked documents does not seem to improve performance substantially, so we land on 25 documents re-ranked to keep computational overhead to a minimum.

Decoding Strategy	Context Block	No Retrieval F1	KF1	B4	RL	RAG DPR-Poly F1	KF1	B4	RL	FiD-RAG DPR-Poly F1	KF1	B4	RL
Beam	No	20.9	17.6	1.7	20.7	23.1	26.5	4.0	24.0	22.8	27.8	4.1	24.1
Beam	Yes	20.6	17.1	1.7	20.4	22.9	25.9	4.1	23.9	22.5	26.7	3.9	23.8
Nucleus: $p = 0.3$	No	20.6	16.0	1.4	20.3	23.0	24.0	3.6	24.2	22.5	23.5	3.5	23.6
Nucleus: $p = 0.3$	Yes	20.1	15.6	1.4	19.9	22.9	23.9	3.7	24.1	22.0	22.9	3.4	23.1
Nucleus: $p = 0.9$	No	17.1	13.6	0.6	17.0	19.3	19.3	1.9	19.8	19.4	20.2	2.3	20.0
Nucleus: $p = 0.9$	Yes	16.6	13.2	0.6	16.8	19.2	18.9	1.8	19.6	19.6	19.8	2.3	20.4
Top-k: $k = 10$	No	18.0	14.4	0.7	18.0	19.8	19.0	1.8	20.3	20.2	19.9	2.2	20.8
Top-k: $k = 10$	Yes	17.5	14.0	0.5	17.5	19.7	18.8	1.8	20.1	19.7	20.2	2.2	20.2

Table 20: **Comparison of Decoding Strategies** For models with and without retrieval-augmentation. Evaluations are conducted on the WoW Valid Seen. Retrieval models are retrieving 5 documents over all of Wikipedia. We set the minimum beam length to 20, and block tri-grams during beam search. All models use BART as the base seq2seq model.

Pre-training Data	Valid Seen PPL	F1	KF1	Valid Unseen PPL	F1	KF1
DPR						
NQ + TQA	11.6	22.5	26.0	13.4	21.8	22.7
WoW	12.1	22.7	26.2	13.4	22.1	24.4
NQ + TQA + WoW	12.1	22.7	25.8	13.7	22.0	23.0
ColBERT						
MS-Marco	12.4	21.8	25.3	13.5	21.9	24.7
WoW	12.6	21.8	26.1	13.6	21.4	24.9
DPR-Poly and Joint DPR/Poly						
WikiTo	11.7	23.0	26.5	13.1	22.6	24.4
NQ + TQA	11.6	23.0	27.4	13.1	22.1	24.7

Table 21: **Comparison between different retriever/re-ranker pre-training schemes** on WoW Valid Seen/Unseen. All models use BART as the base seq2seq model.

Src	Type	Valid Seen PPL	F1	KF1	Valid Unseen PPL	F1	KF1
A	P	11.6	22.5	26.0	13.4	21.8	22.7
B	P	10.9	23.2	27.9	12.4	22.4	23.7
B	S	13.2	22.3	23.9	15.5	21.5	20.1
C	P	**10.7**	**23.3**	**28.3**	**11.7**	**23.0**	**26.3**
C	S	12.8	22.2	24.8	14.4	21.5	21.7

Table 22: **Comparison between using different sources of knowledge** on WoW Valid Seen/Unseen. All models are BART RAG-Token with DPR Retrieval. **A**: All of Wikipedia; **B**: first two paragraphs from all articles in Wikipedia; **C**: first two paragraphs from all articles in Wikipedia covering the WoW dataset. **P**: full passages are used; **S**: sentences are separate passages.

# Docs	Valid Seen PPL	F1	KF1	Valid Unseen PPL	F1	KF1
RAG-Token						
1	12.8	21.9	27.6	15.3	20.5	23.8
5	11.6	22.5	26.0	13.4	21.7	22.7
25	11.6	22.6	24.5	13.0	21.7	21.1
50	11.6	22.4	23.9	13.0	21.8	20.6
RAG-Sequence						
1	12.5	22.1	27.4	14.6	21.1	24.3
5	11.1	21.5	27.9	12.6	20.3	24.6
25	10.6	21.3	27.8	11.4	20.0	24.3
50	10.5	21.2	27.8	11.2	19.9	24.3
RAG-Turn-DTT						
1	12.7	21.3	28.3	15.0	20.1	24.9
5	11.8	21.9	27.7	13.6	21.1	24.3
25	11.7	22.2	26.8	13.2	21.6	23.3
50	11.9	22.2	26.4	13.7	21.7	22.7
RAG-Turn-DO						
1	14.2	22.2	28.1	16.9	21.3	24.7
5	13.3	23.1	26.8	15.5	22.0	23.3
25	13.3	23.1	24.8	15.1	22.2	21.1
50	13.3	22.6	23.7	15.2	22.0	20.0
FiD-RAG						
1	13.0	21.5	28.5	15.5	20.5	23.0
5	11.0	22.9	27.7	12.7	22.0	25.5
25	11.1	22.3	21.2	12.1	22.7	22.3
50	11.7	21.4	18.0	12.6	22.1	19.1
100	12.7	20.4	15.9	13.6	21.4	16.6

Table 23: **Comparison of the effect of conditioning over different numbers of documents at inference time for different models** on WoW Valid Seen/Unseen. All models use a DPR retriever, with BART as the base seq2seq model.