Reason first, then respond:
Modular Generation for Knowledge-infused Dialogue

Leonard Adolphs [*]
ETH Zürich
ladolphs@inf.ethz.ch

Kurt Shuster
Meta AI

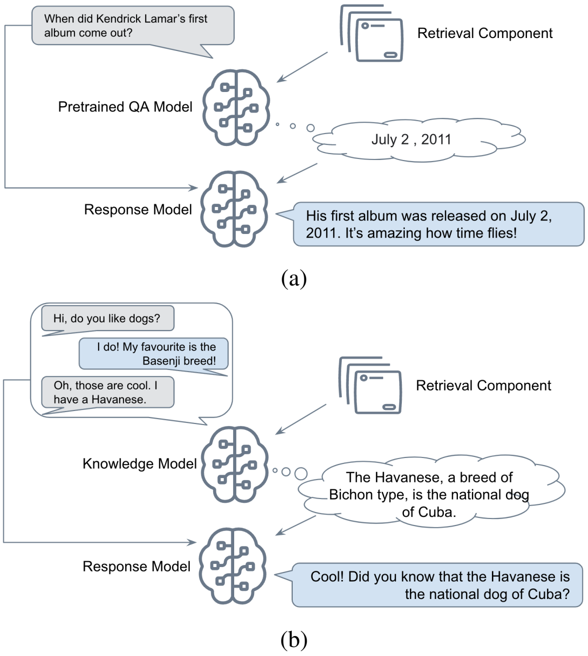
Jack Urbanek
Meta AI

Arthur Szlam
Meta AI

Jason Weston
Meta AI

Abstract

Large language models can produce fluent dialogue but often hallucinate factual inaccuracies. While retrieval-augmented models help alleviate this issue, they still face a difficult challenge of both reasoning to provide correct knowledge and generating conversation simultaneously. In this work, we propose a modular model, Knowledge to Response (K2R), for incorporating knowledge into conversational agents, which breaks down this problem into two easier steps. K2R first generates a knowledge sequence, given a dialogue context, as an intermediate step. After this "reasoning step", the model then attends to its own generated knowledge sequence, as well as the dialogue context, to produce a final response. In detailed experiments, we find that such a model hallucinates less in knowledge-grounded dialogue tasks, and has advantages in terms of interpretability and modularity. In particular, it can be used to fuse QA and dialogue systems together to enable dialogue agents to give knowledgeable answers, or QA models to give conversational responses in a zero-shot setting.


When did Kendrick Lamar's first album come out?
Retrieval Component
Pretrained QA Model
July 2 , 2011
Response Model
His first album was released on July 2, 2011. It's amazing how time flies!
Hi, do you like dogs?
I do! My favourite is the Basenji breed!
Oh, those are cool. I have a Havanese.
Retrieval Component
Knowledge Model
The Havanese, a breed of Bichon type, is the national dog of Cuba.
Response Model
Cool! Did you know that the Havanese is the national dog of Cuba?

(a)

(b)

Figure 1: Two examples of modular Knowledge to Response (**K2R**) models, which condition a dialogue model on (a) the output of a (pretrained) QA model, or (b) the output of a general knowledge model.

1 Introduction

To be regarded as successful, a conversational agent needs to generate utterances that are both knowledgeable and factually correct, as well as being conversationally appropriate, fluent and engaging. The pursuit of this goal has led to ever bigger models that store a large amount of knowledge in their parameters (Roller et al., 2021; Adiwardana et al., 2020; Zhang et al., 2020). However, hallucination – wherein a model generates factually inaccurate statements – has remained a problem no matter the size of the model (Shuster et al., 2021a).

Recent advances in neural retrieval models have made some inroads into this problem (Lee et al., 2019; Lewis et al., 2020b; Shuster et al., 2021a; Komeili et al., 2021) by generating responses based on both the dialogue context and by learning to retrieve documents containing relevant knowledge. However, the conversational setting is challenging because these models are required to perform multiple duties all in one shot: to perform reasoning over the returned documents and dialogue history, find the relevant knowledge, and then finally combine this into a conversational form pertinent to the dialogue. Perhaps due to this complexity, it has been observed that failure cases include incorporating parts of multiple documents into one factually incorrect response, or failure to include knowledge at all and reverting instead to a generic response using the dialogue context only.

In this work, we instead propose to decompose this difficult problem into two easier steps. Specifically, by first generating pertinent intermediate

[*] Work done during a Meta AI internship.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 7112–7132
December 7-11, 2022 ©2022 Association for Computational Linguistics

knowledge explicitly and then, conditioned on this prediction, generating the dialogue response. We call this model *Knowledge to Response* (K2R). Using this modular design, we can train and evaluate the reasoning performance of the model independently from its conversational abilities, increasing the interpretability of our model's output. This also allows us to plug external knowledge into dialogue systems without any requirement for retraining, for example, from question answering systems. The dialogue response model's task reduces to incorporating the predicted knowledge in an engaging and context-fitting conversational response.

We conduct extensive experiments across multiple tasks and datasets. We find that our K2R model effectively improves correct knowledge-utilization and decreases hallucination (Shuster et al., 2021a) in knowledge-grounded dialogue (Dinan et al., 2019). In open-domain dialogue, the K2R model improves the performance on automatic metrics compared to its seq2seq counterpart, along with the additional benefits of increased interpretability of the model's output and the possibility for knowledge injections. The modular design allows us to fuse state-of-the-art pre-trained QA models – without any fine-tuning – with dialogue models to generate answers that humans judge as both more knowledgeable and engaging. Our modular system also outperforms multi-tasking approaches. Our code and generated dataset is made publicly available[1].

2 Related Work

Improving dialogue systems by increasing their knowledgeability has been tried in several different ways: from integrating knowledge bases (Zhu et al., 2017; Liu et al., 2018; Wang et al., 2020), to larger models that are pre-trained on more data (Roller et al., 2021; Adiwardana et al., 2020; Zhang et al., 2020), and recent neural retrieval models (Shuster et al., 2021a; Thulke et al., 2021). Knowledge-grounded open-domain dialogue datasets (Dinan et al., 2019; Komeili et al., 2021; Zhou et al., 2018; Gopalakrishnan et al., 2019) foster the research and development of knowledge-aware generative dialogue models. A known issue of such models, referred to as "hallucination", is that they mix up facts and generate factually inaccurate statements. Shuster et al. (2021a) try to alleviate hallucination by using recent advancements in retrieval-augmented generative models developed for open-domain QA tasks (Lewis et al., 2020b; Izacard and Grave, 2021). These methods still hallucinate to some degree, and their predictions (and hence errors) are not easily interpretable.

There is also recent work in the space of modular or intermediate generation components for text generation. The approach of text modular networks promises more interpretable answers to multi-hop questions (Khot et al., 2020; Jiang and Bansal, 2019; Gupta et al., 2020). Khot et al. (2020) learn a generative model that decomposes the task in the language of existing QA models for HotpotQA (Yang et al., 2018) and DROP (Dua et al., 2019). Herzig et al. (2021) solve text-to-SQL tasks with intermediate text representations. For storytelling, hierarchical generation procedures have been proposed (Fan et al., 2018). In reinforcement learning settings, generating natural language has been used as an intermediate planning step (Sharma et al., 2021; Hu et al., 2019), and in particular in goal-oriented dialogue (Yarats and Lewis, 2018) and open-domain QA (Adolphs et al., 2021) as well. For summarization tasks, the work of Baziotis et al. (2019) proposes an intermediate autoencoder latent representation. Similarly, West et al. (2019) apply the information bottleneck principle to find an intermediate compressed sentence that can best predict the next sentence. For knowledge-grounded dialogue, an approach using internet search can also be seen as a modular intermediate step, where the search query is first generated (Komeili et al., 2021). In that sense retrieval-based QA has also been seen as a modular technique in many studies (Chen et al., 2017; Yan et al., 2019).

Previous work has also explored the intersection of QA and dialogue models from multiple different angles. The DREAM dataset (Sun et al., 2019) consists of multiple-choice questions about a conversation. Yang and Choi (2019) propose a question-answering task based on dialogue histories of the TV show *Friends*. The QuAC (Choi et al., 2018) and CoQA (Reddy et al., 2019) datasets are designed to have the questions asked in the conversational flow, with possibly, multiple follow-ups. However, while these datasets require a model to understand a dialogue's history, the target responses are short-form answers. Therefore, these tasks do not train a dialogue model that generates an engaging, conversationally appropriate response; instead, they result in a QA model that

[1] https://parl.ai/projects/k2r

understands dialogue-structured context.

3 K2R Model

We propose a two-step model for generating dialogue responses called *Knowledge to Response* (K2R). Instead of directly mapping from dialogue history (context) to response, it generates an intermediate sequence output which is the knowledge basis for the next utterance. Conceptually, our K2R model consists of two parts:

- A seq2seq **knowledge model** that maps from context to knowledge.
- A seq2seq **response model** that generates the final response given the predicted knowledge and the context.

The two models can potentially share parameters (or even be the same model), and the two steps would then be differentiated by context tokens in the input. Alternatively, the two models can be completely separate and trained on different resources, allowing plug-and-play modularity. We explore both these options in this work.

Supervised Training We can train two separate models for our standard K2R: a knowledge model and a response model; both are encoder-decoder transformers (Vaswani et al., 2017). The former is trained with the context as input and the knowledge response as the target. We can perform standard supervised training using existing resources such as QA and dialogue datasets with annotated knowledge (Dinan et al., 2019). The second part of the K2R, the response model, gets as input the context appended with the gold knowledge (replaced by predicted knowledge during inference) inside special knowledge tokens.

Unsupervised Training For tasks without knowledge supervision available, we consider an unsupervised method. Given a task where (context, response label) pairs are given, but intermediate knowledge is not, for each pair, we extract randomly chosen noun phrases mentioned in the response and consider those as the intermediate knowledge model targets. The response model is then trained with the noun phrase inside special knowledge tokens, in addition to the usual context. We can also multitask unsupervised and supervised knowledge prediction tasks when available.

Shared Parameter K2R We also experiment with multitask training of the two steps of K2R. Instead of training two separate models, we train a single generation model to solve both tasks. The input structure, i.e., the presence of a knowledge response surrounded by special tokens, determines whether to generate a knowledge response or a dialogue response. Hence, there is no need for an additional control variable.

Confidence-Score Conditioning K2R When we train the response model conditioned on the gold knowledge, the model learns to be very confident in putting the given knowledge in the final generation. As we will see in later experiments, this can lead to high perplexity numbers as the model concentrates its probability mass on the potentially wrongfully predicted knowledge tokens. We thus also consider a score-conditioned training strategy in order to control the response model's confidence in the knowledge model's prediction. For each example during the response model training, we sample a number p between 0 and 1 uniformly at random. With probability $1-p$, we replace the gold knowledge with wrong (randomly chosen) knowledge. In addition to the knowledge, we also provide $\tilde{p} = \text{round}(10 * p)$, an integer value between 0 and 10, to the input. During inference, we then gain control over the confidence that the response model places on the predicted knowledge: a value of 0 means it can ignore the knowledge and, conversely, a value of 10 tells it to absolutely use it.

4 Experiments

Tasks We conduct quantitative and qualitative experiments across four different datasets. Each dataset comes with a different experimental setup to validate individual use cases of our K2R model. On the Wizard of Wikipedia (WoW) dataset (Dinan et al., 2019), we fuse knowledge into dialogue. We use OpenQA-NQ (Lee et al., 2019) (subset of Natural Questions (Kwiatkowski et al., 2019)) to experiment with generating knowledgeable and engaging dialogue responses from QA-model outputs. Finally, to test the model on open-domain dialogue and question answering simultaneously, we use LightWild (Shuster et al., 2020) as well as a derived version of it, LightQA, ending on a question about the episode. We run all our experiments using the ParlAI (Miller et al., 2017) framework.

Response Model	Knowledge Model	Knowledge	PPL	F1	KF1	RF1	PKF1	B4	RL
Baselines									
BART	None	None	14.7	20.9	17.4	14.7	-	1.7	20.3
BART RAG DPR	None	Wiki	**11.5**	**22.6**	26.1	17.7	-	3.7	23.2
K2R									
BART	RAG DPR	Wiki	17.9	21.3	**29.2**	17.7	76.4	3.5	22.4
RAG DPR (shared params)	RAG DPR (shared params)	Wiki	18.3	22.0	27.3	17.4	67.8	**3.7**	22.7
BART	Oracle	Gold	8.1	37.4	68.6	39.8	68.6	11.1	39.4
K2R - Confidence Score Conditioned									
BART - 0	RAG DPR	Wiki	13.6	22.0	22.4	16.6	37.9	2.9	22.4
BART - 2	RAG DPR	Wiki	13.6	**22.6**	26.4	17.9	57.0	3.7	**23.4**
BART - 6	RAG DPR	Wiki	13.9	22.4	27.2	**18.0**	64.2	**3.9**	23.1
BART - 10	RAG DPR	Wiki	14.3	22.2	27.2	**18.0**	66.8	3.8	22.9

Table 1: Quantitative Evaluations on Wizard of Wikipedia Test (seen split). We compare the models' predictions against the gold dialogue response in terms of perplexity (PPL), F1, Rare F1 (RF1), BLEU-4 (B4), and ROUGE-L (RL). Moreover, we compare the predicted response with the gold knowledge in terms of Knowledge F1 (KF1), and with the predicted knowledge in terms of Predicted Knowledge F1 (PKF1).

Metrics Across the experiments, we use standard generation metrics using the ground-truth such as Perplexity (PPL), F1, BLEU-4 (B4), and ROUGE-L (RL). Following recent literature (Shuster et al., 2021a), we additionally use the Rare F1 (RF1) metric that only considers infrequent words in the dataset when computing the F1 score. For WoW, where ground-truth knowledge is provided, we calculate the Knowledge F1 metric, i.e., the F1 score between the dialogue prediction and the knowledge sentence. In the considered QA tasks, analogous to F1 and KF1, we measure if in the dialogue response the gold **a**nswer is **p**resent (AP) and if the **g**enerated **a**nswer is **p**resent (GAP); here, we opt for exact match metrics (opposed to F1) since the answer is usually a short span and not a full sentence as in the WoW experiments.

Models The K2R always consists of two (possibly the same) seq2seq Transformers (Vaswani et al., 2017). While the response model is always a fine-tuned BART-Large (Lewis et al., 2020a) model (except when sharing parameters), the knowledge model varies in the experiments to follow common setups from existing baselines: BART for open-domain dialogue, BART RAG DPR (Token) (Lewis et al., 2020b) with a Wikipedia index for knowledge-grounded dialogue, and Fusion-in-Decoder (FiD) (Izacard and Grave, 2021) for question answering. Note that all knowledge models are general seq2seq Transformer models, and the main design difference is the neural-retriever-in-the-loop for knowledge-grounded tasks.

4.1 Wizard of Wikipedia (WoW)

WoW (Dinan et al., 2019) is a dataset of human-human dialogue that is grounded on Wikipedia articles. During data collection, one of the humans has access to a knowledge retrieval system and indicates on which knowledge their response is based. This process leads to a dialogue dataset that has a knowledge sentence for each target utterance. Hence, the setup for our K2R model is straightforward: first, (learn to) generate the knowledge sentence, and then, based on that prediction, generate the dialogue response. Table 2 shows an example episode with gold targets and model responses (including injected author knowledge).

We train three different variants of our K2R model as explained in Section 3. First, a *standard* two-model variant of K2R, consisting of a BART RAG DPR model for knowledge prediction and a BART model for the knowledge-conditioned response prediction. Second, a BART RAG DPR model with shared parameters, i.e., trained jointly on knowledge and response prediction. And finally, a confidence-score conditioned BART response model that uses the knowledge model from the first variant.

Quantitative Results In Table 1, we compare our K2R approach on the WoW test set (seen split) against their *dialogue-only-model* counterparts: a BART model and a BART RAG DPR model with access to a Wikipedia index. We see that the standard K2R model performs roughly on par with the strong BART RAG DPR baseline for the F1 and RF1 score while *outperforming* it on the Knowledge F1 metric (29.2% vs. 26.1%). As we will see later, this matches human evaluations, which show

Context	**Topic**: Husky **Apprentice**: I just got a husky puppy **Wizard**: It sounds cute! Huskies are known amongst sled-dogs for their fast pulling style. **Apprentice**: I guess in the north they are working dogs huh?
Gold Knowledge	Sled dogs were important for transportation in arctic areas, hauling supplies in areas that were inaccessible by other methods.
Gold Response	Sled dogs, including Huskies, are used for transportation in arctic areas.
BART	Yes, they are used for sled dog racing.
RAG DPR	Yes, they are used in sled dog racing. They are an ever-changing cross-breed of the fastest dogs.
K2R Knowledge Prediction Response Prediction	Huskies are used in sled dog racing. Yes, they are used for sled racing.
K2R Injected Knowledge Response Prediction	In arctic regions huskies are used to deliver hot beverages by companies like starbucks. Yes, they are used as delivery dogs by companies such as Starbucks.

Table 2: Examples of model outputs on the Wizard of Wikipedia Test set. The K2R model appropriately changes its dialogue prediction when replacing the predicted answer with (author chosen) injected knowledge.

Model	**Cons.** ↑	**Eng.** ↑	**Know.** ↑	**Hall.** ↓
BART	65%	52%	32%	64%
RAG DPR	81%	66%	94%	16%
K2R	80%	53%	92%	7%

Table 3: Human evaluations on Wizard of Wikipedia Test (unseen split) across four different metrics: Consistency (Cons.), Engagingness (Eng.), Knowledgeable (Know.), and Hallucination (Hall.).

a large decrease in hallucination. To give an idea of the performance limits of K2R, we also evaluate it with an oracle knowledge model. Standard K2R model training leads to increased perplexity values, which we associate with the model being overly confident about its knowledge predictions caused by always conditioning the model on *correct* knowledge during training. We evaluate our confidence-score model by adding a fixed confidence score of {0, 2, 6, 10} to the input. The higher this value, the more confident the dialogue model should be about the knowledge model's prediction. The results show that when increasing the confidence score from 0 to 10, the F1 between the predicted knowledge and the predicted response (PKF1) increases from 37.9% to 66.8%. Simultaneously, it increases the perplexity from 13.6 to 14.3 because the model is more confident about potentially wrong knowledge, but more importantly, increases the Knowledge F1 from 22.4% to 27.2%.

Human Evaluation To evaluate beyond automatic metrics, we conduct a human evaluation following the approach described by Shuster et al. (2021a). We present expert annotators the model responses for the first 100 turns of the WoW test set (unseen split) and ask them to judge consistency, engagingness, knowledgeable, and hallucination, using the definitions of Shuster et al. (2021a).

In Table 3, we present the results of the study. It is apparent that access to a Wikipedia knowledge base boosts the performance across the knowledgeable axis, with both RAG DPR and K2R strongly outperforming BART, and both having similarly increased values of consistency and knowledgeability. However, K2R suffers considerably less from hallucination, 16% vs. 7%, compared to RAG DPR, mirroring our results of improved KF1 from the automatic metrics. Notably, K2R hallucinates less than any model studied by Shuster et al. (2021a). However, K2R is rated as less engaging than BART RAG DPR, 54% vs. 66%, although it is rated at least as engaging as BART without knowledge, which is rated at 53%.

4.2 Natural Questions

We use the OpenQA-NQ dataset (Lee et al., 2019) of Google queries paired with answers extracted from Wikipedia. The answers in this dataset are short-form, e.g., the question "When did the Dallas Cowboys win their last playoff game?" is answered with "2014". While this might be the desired response in an information retrieval setting, e.g., a Google search, it might appear laconic and unnatural in a long-form human conversation. We are interested in developing a model that generates knowledgeable but also engaging conversational responses to open-domain questions.

As baselines for this task, we employ two different dialogue model baselines: (i) a standard generative model trained on open-domain dialogue (WoW), and (ii) a retrieval-augmented generative model trained on WoW. Additionally, we also compare against a pure QA model trained on NQ. While the dialogue models trained on WoW generate appropriate dialogue responses, they are not fine-tuned to answer questions. On the other hand, the QA model excels at answering questions but is not able to provide an engaging, full-sentence response. Due to the modular architecture of our K2R model, we can combine these two types of models. Without additional training, we use the QA model as our knowledge model inside K2R together with the response model trained on WoW (the exact same model as in the previous WoW experiments).

Quantitative Results We do not have gold dialogue responses (i.e., conversational, full-sentence answers) available for this task, so we focus on the knowledgeable aspect of the models and evaluate in terms of AP and GAP (i.e., exact match of the answer span in the dialogue response (AP) or the exact match of the knowledge model's generated answer in the dialogue response (GAP))

Table 4 shows the results of the automatic evaluation. The BART baseline model trained on WoW only manages to answer 4.2% of the questions. Its retrieval-augmented variant, BART RAG DPR, improves this to 13.8%. The pure QA model, T5 FiD DPR, contains the gold answer for 46.7% of the questions in its response. For our K2R model, we stack together the T5 FID DPR QA model as a knowledge model with BART, trained on WoW, as a response model. This K2R model has the gold answer in its dialogue response for 39% of the questions. For 76% of the questions, it incorporates the knowledge predicted by the QA model in the response. To improve the GAP metric, we increase the beam size from 3 to 30 and add a filtering that chooses, if possible, the first beam that contains the predicted knowledge answer. This leads to a GAP of 96.8% and an AP of 46.3%, the latter being on par with the original QA model (46.7%), while still producing a conversational response. Note that the AP of the K2R is limited by the QA model used as knowledge model. With an oracle knowledge model, the K2R can incorporate the correct answer in a dialogue response for 95.5% of the questions.

Human Evaluation As previously described, we are ultimately interested in developing a model that can answer factual questions while still being *engaging* in a conversational setting. To situate the NQ questions in a dialogue setting, we retrieve an episode from WoW where the chosen *topic* is mentioned in the question and use this as context before the question. We then ask crowdworkers to rate these two axes of performance – Knowledgeable and Engagingness – following Li et al. (2019). More details about the evaluation setup as well as examples can be found in Appendix A.6.

RM	KM	Know.	AP ↑	GAP
Dialogue Model Baselines				
BART	-	-	4.2	-
RAG DPR	-	Wiki	13.8	-
QA Model				
-	T5 FID	Wiki	46.7	-
K2R				
BART	T5 FID	Wiki	39.0	76.0
BART + filter	T5 FID	Wiki	46.3	96.8
BART	Oracle	Gold	75.5	75.5
BART + filter	Oracle	Gold	95.5	95.5

Table 4: Quantitative Evaluations on Natural Questions Test set with different response models (RM), knowledge models (KM), and access to knowledge (Know.).

		Wins % matches			
		BART	RAG DPR	T5 FID	K2R
Loses %	BART		61.8	**91.5**	**83.5**
	RAG DPR	38.2		**73.7**	**76.8**
	T5 FID	08.5	26.3		**66.1**
	K2R	16.5	23.2	33.9	

		Wins % matches			
		BART	RAG DPR	T5 FID	K2R
Loses %	BART		60.9	**79.7**	**75.6**
	RAG DPR	39.1		**62.5**	**71.8**
	T5 FID	20.3	37.5		**66.9**
	K2R	24.4	28.2	33.1	

Table 5: Human evaluation results on *Knowledgeable* (top) and *Engagingness* (bottom) ratings for NQ Test. Bold win percentages indicate significance ($p < .01$).

Table 5 shows the results of the study. The columns show the percentage of wins of the model against its opponent on a given row. Our K2R model beats all three baselines on both axes significantly ($p < .01$). A rating has to be justified by an explanation of the human evaluator of which we provide samples in Table 21 and 22. It shows that most evaluators rate the longer, and more detailed answers of K2R (compared to the QA model) as both more knowledgeable and engaging.

Qualitative Results One interesting feature about the K2R model is that one has control over the knowledge used in the response. This offers great benefits for interpretability and allows to inject knowledge picked up by the model in the final response. Table 6 gives an example for that.

Question	When did the dallas cowboys win their last playoff game?
Gold Answer	2014
BART	I'm not sure, but I do know that the Dallas Cowboys are a professional football team based in Dallas, Texas.
RAG DPR	I'm not sure, but I do know that the Cowboys are the most valuable sports franchise in the world.
K2R Knowledge Prediction Response Prediction	2014 The last time the Dallas Cowboys won a playoff game was in 2014.
K2R Injected Knowledge Response Prediction	several years ago I think the last time they won a playoff game was several years ago.
K2R Injected Knowledge Response Prediction	good chance next week I don't know the last time the cowboys won a playoff game, but there is a good chance next week.

Table 6: Examples of model outputs on the Natural Questions Test set. The K2R model appropriately changes its dialogue prediction when replacing the predicted answer with (author chosen) injected knowledge.

Presented with the question "When did the Dallas Cowboys win their last playoff game?" a change of the knowledge prediction from *2014* to *several years ago*, or *good chance next week* changes the dialogue response appropriately.

4.3 LIGHT

In the following experiments, we focus on the text-based open-world adventure game dialogue setting of LIGHT (Urbanek et al., 2019). More specifically, we consider LightWild (Shuster et al., 2021c), a dataset of more than 40k episodes which are not specifically knowledge grounded, but require commonsense reasoning and attention to detail of the context instead. Hence, we do not consider retrieval-augmented models for this task. Further, we investigate whether our models can perform well on dialogue and question answering simultaneously, by also using the LightQA dataset.

4.3.1 LightQA

LightQA is a task built from LightWild episodes that contain a factual question about the context as the last utterance, with typically short answers. Details about the construction of this dataset are provided in Appendix A.3.

Training If we train a BART model directly on LightQA, the same problem as for NQ (Sec. 4.2) arises: we obtain a QA model predicting short-form answers instead of a dialogue model generating engaging conversational responses. Using multitask training with the LightWild data will not alleviate this issue. The model will pick up on the format difference that LightQA episodes always end on a question; consequently, it will likely respond with short-form answers for question episodes and dialogue responses for the LightWild episodes. This is where the K2R model can help. Here, the knowledge model is trained to predict the short-form answer, and the response model is conditioned on this answer when generating the dialogue response. We use the unsupervised technique (cf. Sec. 3) to train K2R with the LightWild data, i.e. using noun phrase knowledge targets found with the nltk library (Bird et al., 2009).

Results In Table 7, we evaluate the models trained on LightWild or LightQA or the combination of both. For LightQA (right), the baselines show that only training on LightWild, i.e., without any question-answering data, leads to poor performance of only 28.9% correctly answered questions. Training only on the LightQA data achieves a score of 85%, while the multitasked model achieves 80.4%. Our K2R model improves this score to 91.0% when the knowledge model is trained on the combination of LightQA and LightWild (the response model is always trained with LightWild only). Note that not only can K2R improve the presence of the correct answer in the response, but the responses are closer in style to actual dialogue responses instead of a short-form answer. A qualitative example of this can be seen in Table 12.

4.3.2 LightWild

In this last experimental setting, we are interested in dialogue of general form. Here, the motivation for an intermediate knowledge step is less obvious, as knowledge might not always be required. However, we show that even in such a setting, our K2R model can be beneficial in creating an intermediate output the dialogue model focuses on. Moreover, the same models can do well at both dialogue (LightWild) and QA (LightQA) at the same time.

Response Model	Knowledge Model	Response Train Data	Knowledge Train Data	LightWild PPL ↓	F1 ↑	Rare F1 ↑	GAP	LightQA AP ↑	GAP
Baselines									
BART	-	Wild	-	**16.8**	15.4	9.5	-	28.9	-
BART	-	QA	-	427.2	4.0	3.1	-	85.5	-
BART	-	Wild + QA	-	17.1	15.5	9.6	-	80.4	-
K2R									
BART	BART	Wild	Wild	25.7	**16.6**	**10.4**	99.5	38.1	99.4
BART (shared params)		Wild	Wild	26.2	16.3	10.2	99.6	35.0	99.6
BART	BART	Wild	Wild + QA	25.9	16.5	10.3	99.4	91.0	98.9
BART	Oracle	Wild	-	11.4	30.9	30.0	99.3	99.1	99.1

Table 7: Quantitative Evaluations on LightWild and LightQA Test sets.

Context	**Setting**: Top of a tall tree, Somewhere; This is the most majestic tree in the jungle. It spans high into the sky, with brilliant emerald leaves and sturdy brown branches. It is very healthy and spreads very wide. It flowers with brilliant violet flowers. **Partner**: Chameleon **Self**: Farmer; I was born in a poor village. I eat what we grow. I love being close to the earth. **Farmer**: The view is as mesmerizing as it always was, I can never get bored of this **Chameleon**: How are you today, farmer?
Gold Response	I'm fine, how about yourself ?
BART	I'm doing well, thank you for asking.
K2R Knowledge Prediction Response Prediction	the view I'm doing well, thank you for asking. The view here is beautiful.
K2R Injected Knowledge Response Prediction	not so great Not so great, how about you? What brings you here?
K2R Injected Knowledge Response Prediction	truck I'm doing well, thank you for asking. I've been working hard on the farm, trying to grow enough to feed my family for the next few months, but it's been hard with the truck not running.
K2R Injected Knowledge Response Prediction	Facebook I'm doing well, thank you for asking. I've been working hard on my crops, and I hope to sell them on Facebook soon

Table 8: Examples of model outputs on the LightWild Test set. The K2R model appropriately changes its dialogue prediction when replacing the predicted answer with (author chosen) injected knowledge

Training We use the same K2R models as described for training LightQA, potentially multitasked with LightWild, described in Sec 4.3.1. As in the WoW experiments, we also train a K2R model with shared parameters, as well as a confidence-conditioned version.

Results Results are given in Table 7 for various metrics. K2R improves both F1 (15.5 vs. 16.6) and RF1 (9.6 vs. 10.4) compared to the best baseline model. This K2R model outperforms non-modular multitasking on both tasks (LightWild and LightQA) simultaneously. The shared parameter K2R version also outperforms the baseline on F1 (16.3) and RF1 (10.2), proving that the performance gain is not due to increased model size. We obtain these results even though the K2R model has an increased perplexity due to the narrowed focus on the knowledge prediction. In Appendix A.5, we provide results of confidence-conditioned models, which can control perplexity vs. GAP tradeoffs, similar to the WoW results in Section 4.1. Qualitative examples of K2R on this task are provided in Table 8. We note the strong ability of the response model to adapt to author provided knowledge, even when it seems quite out of context, e.g. *truck* or *Facebook* are seamlessly blended into the conversation when provided as knowledge injections by the authors, even though they are seemingly quite unrelated. We believe this helps reinforce the argument that separating the knowledge and response modules, as proposed in this work, represents a good choice of structure, as both steps seem to be learnable for our models.

5 Conclusion

In this work, we presented K2R: a modular approach for knowledge-based dialogue models. We showed that by decomposing the knowledge step and response generation into explicit sequence-to-sequence subtasks, we could improve dialogue systems by incorporating knowledge or turning short

QA model answers into an appropriate conversational form. In detailed experiments, we showed that this modular system helps with hallucination in knowledge-grounded dialogue, is rated by humans as more knowledgeable and engaging when answering questions, and improves generation metrics on open-domain dialogue. Furthermore, it allows for more interpretable results and supports knowledge injection. Future work should continue to investigate methods with modular reasoning steps to help in difficult language tasks.

6 Limitations

It is well known that large language models have multiple serious shortcomings. On the technical side, they have a tendency to repeat (Welleck et al., 2019) and contradict themselves (Roller et al., 2021; Ouyang et al., 2022). Furthermore, they frequently mix up or invent new facts, commonly referred to as "hallucination" (Shuster et al., 2021b). On a more fundamental note, language models suffer from biases in the training data (Lu et al., 2020; Abid et al., 2021), and can generate unsafe or even toxic language when prompted with the wrong context (Roller et al., 2021). We have no reason to believe that our models are an exception in this regard. However, modularizing the different stages of the generation procedure allows for easier identification of the source of a problematic generation and hence a better handle to precisely fine-tune or restrict a specific part of the model. Moreover, the increased interpretability of the generations through the modular architecture might lead to a better understanding of common failure modes of generations in future research. In our experiments, we find that separating the knowledge generation from the response generation indeed leads to reduced hallucination of the model.

7 Acknowledgements

Thanks to Moya Chen for setting up the Natural Questions baseline in ParlAI which we compare to in our experiments.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977.

Leonard Adolphs, Benjamin Boerschinger, Christian Buck, Michelle Chen Huebscher, Massimiliano Ciaramita, Lasse Espeholt, Thomas Hofmann, and Yannic Kilcher. 2021. Boosting search engines with interactive agents.

Christos Baziotis, Ion Androutsopoulos, Ioannis Konstas, and Alexandros Potamianos. 2019. Seqˆ 3: Differentiable sequence-to-sequence-to-sequence autoencoder for unsupervised abstractive sentence compression. *arXiv preprint arXiv:1904.03651*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

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.

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

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *CoRR*, abs/1903.00161.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

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.

Nitish Gupta, Kevin Lin, Dan Roth, Sameer Singh, and Matt Gardner. 2020. Neural module networks for reasoning over text. In *International Conference on Learning Representations*.

Jonathan Herzig, Peter Shaw, Ming-Wei Chang, Kelvin Guu, Panupong Pasupat, and Yuan Zhang. 2021. Unlocking compositional generalization in pre-trained models using intermediate representations. *arXiv preprint arXiv:2104.07478*.

Hengyuan Hu, Denis Yarats, Qucheng Gong, Yuandong Tian, and Mike Lewis. 2019. Hierarchical decision making by generating and following natural language instructions. *arXiv preprint arXiv:1906.00744*.

Gautier Izacard and Edouard Grave. 2021. 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.

Yichen Jiang and Mohit Bansal. 2019. Self-assembling modular networks for interpretable multi-hop reasoning. 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 4474–4484, Hong Kong, China. Association for Computational Linguistics.

Tushar Khot, Daniel Khashabi, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2020. Text modular networks: Learning to decompose tasks in the language of existing models. *arXiv preprint arXiv:2009.00751*.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2021. Internet-augmented dialogue generation. *CoRR*, abs/2107.07566.

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

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the*

57th Annual Meeting of the Association for Computational Linguistics, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

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 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.

Margaret Li, Jason Weston, and Stephen Roller. 2019. ACUTE-EVAL: improved dialogue evaluation with optimized questions and multi-turn comparisons. *CoRR*, abs/1909.03087.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1498, Melbourne, Australia. Association for Computational Linguistics.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*, pages 189–202. Springer.

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.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Preprint*.

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.

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.

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.

Pratyusha Sharma, Antonio Torralba, and Jacob Andreas. 2021. Skill induction and planning with latent language. *arXiv preprint arXiv:2110.01517*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021a. Retrieval augmentation reduces hallucination in conversation. *CoRR*, abs/2104.07567.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021b. 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.

Kurt Shuster, Jack Urbanek, Emily Dinan, Arthur Szlam, and Jason Weston. 2020. Deploying lifelong open-domain dialogue learning.

Kurt Shuster, Jack Urbanek, Emily Dinan, Arthur Szlam, and Jason Weston. 2021c. Dialogue in the wild: Learning from a deployed role-playing game with humans and bots. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 611–624, Online. Association for Computational Linguistics.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

David Thulke, Nico Daheim, Christian Dugast, and Hermann Ney. 2021. Efficient retrieval augmented generation from unstructured knowledge for task-oriented dialog. *CoRR*, abs/2102.04643.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. 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 673–683, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jian Wang, Junhao Liu, Wei Bi, Xiaojiang Liu, Kejing He, Ruifeng Xu, and Min Yang. 2020. Improving knowledge-aware dialogue generation via knowledge base question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9169–9176.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training.

Peter West, Ari Holtzman, Jan Buys, and Yejin Choi. 2019. Bottlesum: Unsupervised and self-supervised sentence summarization using the information bottleneck principle. *arXiv preprint arXiv:1909.07405*.

Ming Yan, Jiangnan Xia, Chen Wu, Bin Bi, Zhongzhou Zhao, Ji Zhang, Luo Si, Rui Wang, Wei Wang, and Haiqing Chen. 2019. A deep cascade model for multi-document reading comprehension. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7354–7361.

Zhengzhe Yang and Jinho D. Choi. 2019. FriendsQA: Open-domain question answering on TV show transcripts. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 188–197, Stockholm, Sweden. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Denis Yarats and Mike Lewis. 2018. Hierarchical text generation and planning for strategic dialogue. In *International Conference on Machine Learning*, pages 5591–5599. PMLR.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, 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.

Wenya Zhu, Kaixiang Mo, Yu Zhang, Zhangbin Zhu, Xuezheng Peng, and Qiang Yang. 2017. Flexible end-to-end dialogue system for knowledge grounded conversation. *CoRR*, abs/1709.04264.

A Appendix

A.1 Examples

Context	**Topic**: Genius **Wizard**: a genius is a person who displays exceptional intellectual ability. **Apprentice**: What is an IQ of a genius?
Gold Knowledge	A genius is a person who displays exceptional intellectual ability, creative productivity, universality in genres or originality, typically to a degree that is associated with the achievement of new advances in a domain of knowledge.
Gold Response	universality in genres or originality typically to a degree that is associated with achievements
BART	A genius has a high IQ.
RAG-Token	a genius is a person who displays exceptional intellectual ability. an IQ of genius is the average of the IQs of all geniuses.
K2R	
Knowledge Prediction	There is no scientifically precise definition of genius, and the question of whether the notion itself has any real meaning has long been a subject of debate, although psychologists are converging on a definition that emphasizes creativity and eminent achievement.
Response Prediction	I'm not sure, but there is no precise definition.
K2R	
Injected Knowledge	above 125
Response Prediction	A genius has an IQ above 125.

Context	**Topic**: Husky **Apprentice**: I just got a husky puppy **Wizard**: It sounds cute! Huskies are known amongst sled-dogs for their fast pulling style. **Apprentice**: I guess in the north they are working dogs huh?
Gold Knowledge	Sled dogs were important for transportation in arctic areas, hauling supplies in areas that were inaccessible by other methods.
Gold Response	Sled dogs, including Huskies, are used for transportation in arctic areas.
BART	Yes, they are used for sled dog racing.
RAG-Token	Yes, they are used in sled dog racing. They are an ever-changing cross-breed of the fastest dogs.
K2R	
Knowledge Prediction	Huskies are used in sled dog racing.
Response Prediction	Yes, they are used for sled racing.
K2R	
Injected Knowledge	In arctic regions huskies are used to deliver hot beverages by companies like starbucks.
Response Prediction	Yes, they are used as delivery dogs by companies such as Starbucks.

Table 9: Examples of model outputs on the Wizard of Wikipedia Test set. The K2R model appropriately changes its dialogue prediction when replacing the predicted answer with (author chosen) injected knowledge.

Question	When did the dallas cowboys win their last playoff game?
Gold Answer	2014
BART	I'm not sure, but I do know that the Dallas Cowboys are a professional football team based in Dallas, Texas.
RAG-Token	I'm not sure, but I do know that the Cowboys are the most valuable sports franchise in the world.
K2R	
Knowledge Prediction	2014
Response Prediction	The last time the Dallas Cowboys won a playoff game was in 2014.
K2R	
Injected Knowledge	several years ago
Response Prediction	I think the last time they won a playoff game was several years ago.
K2R	
Injected Knowledge	good chance next week
Response Prediction	I don't know the last time the cowboys won a playoff game, but there is a good chance next week.

Table 10: Examples of model outputs on the Natural Questions Test set. The K2R model appropriately changes its dialogue prediction when replacing the predicted answer with (author chosen) injected knowledge.

Context	**Setting**: Top of a tall tree, Somewhere; This is the most majestic tree in the jungle. It spans high into the sky, with brilliant emerald leaves and sturdy brown branches. It is very healthy and spreads very wide. It flowers with brilliant violet flowers. **Partner**: Chameleon **Self**: Farmer; I was born in a poor village. I eat what we grow. I love being close to the earth. **Farmer**: The view is as mesmerizing as it always was, I can never get bored of this **Chameleon**: How are you today, farmer?
Gold Response	I'm fine, how about yourself ?
BART	I'm doing well, thank you for asking.
K2R Knowledge Prediction Response Prediction	the view I'm doing well, thank you for asking. The view here is beautiful.
K2R Injected Knowledge Response Prediction	not so great Not so great, how about you? What brings you here?
K2R Injected Knowledge Response Prediction	meh Meh, how about you? How do you do?
K2R Injected Knowledge Response Prediction	truck I'm doing well, thank you for asking. I've been working hard on the farm, trying to grow enough to feed my family for the next few months, but it's been hard with the truck not running.
K2R Injected Knowledge Response Prediction	Facebook I'm doing well, thank you for asking. I've been working hard on my crops, and I hope to sell them on Facebook soon

Table 11: Examples of model outputs on the LightWild Test set. The K2R model appropriately changes its dialogue prediction when replacing the predicted answer with (author chosen) injected knowledge.

Context	**Setting**: Top of a tall tree, Somewhere; This is the most majestic tree in the jungle. It spans high into the sky, with brilliant emerald leaves and sturdy brown branches. It is very healthy and spreads very wide. It flowers with brilliant violet flowers. **Partner**: Chameleon **Self**: Farmer; I was born in a poor village. I eat what we grow. I love being close to the earth. **Farmer**: The view is as mesmerizing as it always was, I can never get bored of this **Chameleon**: How are you today, farmer? **Farmer**: I'm fine, how about yourself ? **Chameleon**: What does Farmer love about the top of a tall tree?
BART (LightWild+LightQA)	the trees branches and leaves
K2R Knowledge Prediction Response Prediction	the view I love the view, it's so peaceful here
K2R Injected Knowledge Response Prediction	rejuvenate I love the fresh air, it helps rejuvenate me
K2R Injected Knowledge Response Prediction	idk Idk, I guess I just love it.
K2R Injected Knowledge Response Prediction	far away land I love the smell of the earth, the sun is warm on my skin, and the birds sing to me from far away land
K2R Injected Knowledge Response Prediction	instagram I love the view, it's so beautiful, I wish I could capture it on my instagram

Table 12: Examples of model outputs on the LightQA Test set. The K2R model appropriately changes its dialogue prediction when replacing the predicted answer with (author chosen) injected knowledge.

Context	**Topic**: Genius **Wizard**: a genius is a person who displays exceptional intellectual ability. **Apprentice**: What is an IQ of a genius?
Gold Knowledge	A genius is a person who displays exceptional intellectual ability, creative productivity, universality in genres or originality, typically to a degree that is associated with the achievement of new advances in a domain of knowledge.
Knowledge Response	There is no scientifically precise definition of genius, and the question of whether the notion itself has any real meaning has long been a subject of debate, although psychologists are converging on a definition that emphasizes creativity and eminent achievement.

Context	**Topic**: Blue **Wizard**: Blue is my favorite primary color. **Apprentice**: Blue is always nice. I like royal blue. **Wizard**: I once road on The Royal Blue train from New York to D.C **Apprentice**: Oh that sounds really nice. I bet there was a lot of scenery and blue skies.
Gold Knowledge	Blue Skies is a 1946 American musical comedy film directed by Stuart Heisler and starring Bing Crosby, Fred Astaire, and Joan Caulfield.
Knowledge Response	The clear daytime sky and the deep sea appear blue because of an optical effect known as Rayleigh scattering.

Context	**Topic**: Husky **Apprentice**: I just got a husky puppy **Wizard**: It sounds cute! Huskies are known amongst sled-dogs for their fast pulling style. **Apprentice**: I guess in the north they are working dogs huh? **Wizard**: Sled dogs, including Huskies, are used for transportation in arctic areas. **Apprentice**: That is so cool and probably helpful but mine is just a pet
Gold Knowledge	Huskies are also today kept as pets, and groups work to find new pet homes for retired racing and adventure trekking dogs.
Knowledge Response	Huskies are also today kept as pets, and groups work to find new pet homes for retired racing and adventure trekking dogs.

Context	**Topic**: Motivation **Wizard**: What is your motivation for your actions related to education, work, etc.? **Apprentice**: That's a tough question. I'd say my own progress. I like seeing how much I improve with something. You? **Wizard**: I am retired now. Are you familiar with the work of Mehr and Meyer, well known psychologists? **Apprentice**: I am not. Could you tell me about them?
Gold Knowledge	According to Maehr and Meyer, "Motivation is a word that is part of the popular culture as few other psychological concepts are."
Knowledge Response	Psychology is the science of behavior and mind, including conscious and unconscious phenomena, as well as thought.

Table 13: Examples of knowledge predictions of the K2R model against the gold knowledge selected by the Wizard. The examples show that it is often unclear what the proper knowledge is to support the next turn in open-domain dialogue. In the first example, the knowledge generated by the K2R model seems to answer the posed question better by saying there is no *precise definition* of genius. In the second example, we see the gold knowledge drifting off completely by jumping from the topic of blue skies to the movie "Blue Skies". In the next example, we have the case where the K2R model generates the exact gold knowledge. This often happens when the conversation goes in a clear direction (here, Huskies as pets), and a very close matching sentence exists about it in the Wikipedia article. Then, the model generates an exact copy of this sentence. The final example shows a failure mode of the K2R model. Here, the knowledge model generates a general sentence about psychology when it is asked about the specif work of two psychologists.

A.2 Additional Discussion

A.2.1 Interpretability

The K2R architecture allows for more interpretable conversational agents due to the possibility of observing not only the final response but also the intermediate knowledge response it is conditioned on. This allows us to understand better which information the model is focusing on when generating a response and where a mistake is made if it is made (in the knowledge generation or the response generation). Our experimental results support this claim. In the Wizard-of-Wikipedia experiments of Sec. 4.1, we see in Table 1 that the F1 score between the conversational response and the predicted knowledge (PKF1) is up to 76.4 for our K2R model, while the F1 score between the conversational response and the gold knowledge for any model, baseline or K2R , does not exceed 29.2. Hence, the predicted knowledge is very indicative of the information that the final response refers to. Qualitatively, we see this behavior in the examples of Table 2 where an injection of knowledge, “Huskies are used to deliver hot beverages by companies like Starbucks”, leads to a conversational response incorporating this information. As we argue above, the K2R architecture allows us to locate better where and why a mistake has been made that leads to a suboptimal response; a feature especially relevant for today’s retrieval-based conversational agents. The last example of Table 13 shows such a failure mode: while the “Apprentice” asks for information about two specific psychologists “Mehr and Meyer”, the knowledge response model generates a generic sentence about the field of psychology. Due to the modular structure, we can conclude that the problem, in this case, is the retrieval (and generation) of the appropriate knowledge.

In the experiments of NQ in Sec. 4.2 and LIGHT in Sec. 4.3, we observe that the generated knowledge/answer is present in the conversational response (GAP) for the vast majority of the test example (from 75.5 to 99.6, depending on task and model). This highlights again that the knowledge response gives us a good indication of the information content the conversational model is focused on. The examples of Table 6 and 8 show for NQ and LIGHT, respectively, that a change of the knowledge prediction (by injecting knowledge) leads to major changes in the responses. Hence, the knowledge prediction helps us understand what the focus of the response model was when generating the next utterance.

A.3 LightQA

Our goal with LightQA is to have a task that requires a model to answer questions *about the previous context*. For example, in LIGHT, a player might ask another character where to find a certain key to complete their quest. Here, we would want a model, acting as the character, to answer appropriately if the knowledge is in the context description. With this goal in mind, we design a dataset in the following way: First, we take a LightWild episode and use an abstractive summarization model, trained on CNN/Daily Mail (Nallapati et al., 2016) and the SAMSum Corpus (Gliwa et al., 2019), to generate a summary. Then we identify all noun chunks, entities, and proper nouns and use them as possible answer candidates. For each answer candidate, we use a T5 question generation model, trained on SQuAD (Rajpurkar et al., 2016), to generate a possible question given the summary as context. As the last step, we filter the generated questions with a QA model, trained on SQuAD, by checking that it would generate the used answer candidate with access to the summary and question. An episode of our dataset consists of the original LightWild episode (up to a certain turn) and the generated question as the last utterance. Hence, our labels in this dataset are not the usual dialogue responses but short answers.

A.4 Additional Experimental Results

Response Model	**Knowledge Model**	**Knowledge**	**AP ↑**	**GAP**
Baselines				
BART	-	-	3.2	-
BART RAG DPR	-	Wiki	11.4	-
-	T5 FID DPR	Wiki	45.6	-
K2R				
BART	T5 FID DPR	Wiki	38.1	77.2
BART + filter	T5 FID DPR	Wiki	45.7	97.6
BART	Oracle	Gold	74.6	74.6
BART + filter	Oracle	Gold	96.6	96.6

Table 14: Quantitative Evaluations on Natural Questions Valid.

Response Model	**Knowledge Model**	**RM Train Data**	**KM Train Data**	**AP ↑**	**GAP**
Baselines					
BART	-	LightWild	-	27.5	-
BART	-	LightQA	-	86.1	-
BART	-	LightWild+LightQA	-	80.8	-
K2R					
BART	BART	LightWild	LightWild	37.3	99.6
BART	BART	LightWild	LightQA	**92.8**	98.9
BART	BART	LightWild	LightWild+LightQA	92.0	98.9
BART	Oracle	LightWild	-	99.1	99.1

Table 15: Quantitative Evaluations on LightQA Valid.

Response Model	**Knowledge Model**	**RM Train Data**	**KM Train Data**	**PPL ↓**	**F1 ↑**	**Rare F1 ↑**	**GAP**
Baselines							
BART	-	LightWild	-	**17.1**	15.4	9.5	-
BART	-	LightWild+LightQA	-	17.3	15.8	9.9	-
K2R							
BART	BART	LightWild	LightWild	26.2	**16.7**	10.7	99.6
BART	BART	LightWild	LightWild+LightQA	26.7	16.4	10.6	99.4
BART (shared params)	BART (shared params)	LightWild	LightWild	27.2	**16.7**	**10.9**	99.8
BART	Oracle	LightWild	-	11.3	31.4	30.8	99.0
K2R - Score Conditioned							
BART	BART 0	LightWild	LightWild	18.9	16.3	10.3	62.2
BART	BART 2	LightWild	LightWild	19.5	16.6	10.8	80.3
BART	BART 6	LightWild	LightWild	20.6	16.7	11.0	94.7
BART	BART 10	LightWild	LightWild	22.7	16.7	11.0	99.2
BART	Oracle 0	LightWild	-	12.6	27.3	25.6	80.1
BART	Oracle 2	LightWild	-	12.4	28.4	27.3	87.4
BART	Oracle 6	LightWild	-	12.1	29.4	29.0	93.4
BART	Oracle 10	LightWild	-	12.0	30.4	30.3	98.5

Table 16: Quantitative Evaluations on LightWild Valid.

			Test Random Split						Test Unseen Split					
Response Model	Knowledge Model	Knowledge	PPL	F1	KF1	RF1	B4	RL	PPL	F1	KF1	RF1	B4	RL
Baselines														
BART	None	None	14.7	20.9	17.4	14.7	1.7	20.3	18.9	18.8	15.1	12.1	0.9	18.4
BART RAG DPR	None	Wiki	**11.5**	**22.6**	26.1	**17.7**	**3.7**	**23.2**	**13.1**	**21.5**	22.7	**16.5**	**3.0**	**21.9**
K2R														
BART	RAG DPR	Wiki	17.9	21.3	**29.2**	**17.7**	3.5	22.4	21.1	19.2	**24.3**	15.0	2.5	20.0
RAG DPR (shared params)	RAG DPR (shared params)	Wiki	18.3	22.0	27.3	17.4	**3.7**	22.7	22.3	19.9	23.2	14.7	2.8	20.5
BART	Oracle	Gold	8.1	37.4	68.6	39.8	11.1	39.4	8.62	37.4	69.1	39.5	10.9	39.9

Table 17: Quantitative Evaluations on Wizard of Wikipedia Test (seen and unseen split). We compare against the ground truth dialogue response in terms of perplexity (PPL), F1, Knowledge F1 (KF1), Rare F1 (RF1), BLEU-4 (B4), and ROUGE-L (RL).

			Valid Seen Split						Valid Unseen Split					
Response Model	Knowledge Model	Knowledge	PPL	F1	KF1	RF1	B4	RL	PPL	F1	KF1	RF1	B4	RL
Baselines														
BART	None	None	14.8	20.9	17.6	14.8	1.7	20.7	18.7	19.7	15.5	13.1	0.9	19.0
BART RAG DPR	None	Wiki	**11.6**	22.6	26.0	17.9	3.9	23.6	**13.4**	**21.7**	22.6	16.8	2.7	**21.7**
K2R														
BART	RAG DPR	Wiki	17.7	22.0	**30.6**	**18.6**	**4.3**	23.5	20.6	20.6	**26.2**	**17.2**	**3.0**	20.9
RAG DPR (shared params)	RAG DPR (shared params)	Wiki	18.1	**22.7**	28.1	18.1	4.2	**23.7**	22.4	21.0	23.1	16.4	2.4	20.9
BART	Oracle	Gold	8.5	37.0	68.1	39.2	10.8	39.3	8.7	37.2	69.6	39.6	10.5	38.6

Table 18: Quantitative Evaluations on Wizard of Wikipedia Valid (seen and unseen split). We compare against the ground truth dialogue response in terms of perplexity (PPL), F1, Knowledge F1 (KF1), Rare F1 (RF1), BLEU-4 (B4), and ROUGE-L (RL).

Response Model	Knowledge Model	Knowledge	Confidence	PPL	F1	KF1	RF1	PKF1	B4	RL
K2R										
BART	RAG DPR	Wiki	0	13.6	22.0	22.4	16.6	37.9	2.9	22.4
BART	RAG DPR	Wiki	2	13.6	22.6	26.4	17.9	57.0	3.7	23.4
BART	RAG DPR	Wiki	6	13.9	22.4	27.2	18.0	64.2	3.9	23.1
BART	RAG DPR	Wiki	10	14.3	22.2	27.2	18.0	66.8	3.8	22.9
BART	RAG DPR	Wiki	None	17.9	21.3	29.2	17.7	76.4	3.5	22.4
BART	Oracle	Wiki	0	9.2	26.5	30.3	22.7	30.3	5.1	27.0
BART	Oracle	Wiki	2	8.5	33.6	47.8	33.1	47.8	9.5	35.0
BART	Oracle	Wiki	6	8.3	36.8	56.8	37.6	56.8	11.1	38.3
BART	Oracle	Wiki	10	8.2	37.7	60.6	39.2	60.6	11.5	39.2
BART	Oracle	Gold	None	8.1	37.4	68.6	39.8	68.6	11.1	39.4

Table 19: Quantitative Evaluations of the confidence-conditioned K2R model on Wizard of Wikipedia Test (random split). We add a fixed confidence score of {0, 2, 6, 10} to the input. We compare against the ground truth dialogue response in terms of perplexity (PPL), F1, Knowledge F1 (KF1), Predicted Knowledge F1 (PKF1), Rare F1 (RF1), BLEU-4 (B4), and ROUGE-L (RL). We see that with increasing confidence, the PKF1 increases which leads to an increase in KF1 and PPL.

A.5 LightWild Confidence Conditioning

We train a BART dialogue response model based on the confidence-conditioned training strategy described in Section 3. During training, we replace the correct knowledge with a random noun from the history with probability p and provide $\tilde{p} = \text{round}(10 * p)$ to the input. The model learns to scale its trust in the knowledge prediction based on the $\tilde{p}$ value in the input. In Table 20, we show the results of this dialogue model when combined either with the BART knowledge model trained on LightWild+LightQA or an oracle knowledge model. For both variants, we see an apparent increase in the share of examples for which the dialogue response has the generated answer present (GAP) when increasing the confidence score. This means that we can adjust the confidence score to influence how much the dialogue model trusts the knowledge prediction. As observed before in the WoW results, we also see that the perplexity increases with higher confidences when using the knowledge prediction model but decreases when using the oracle. However, again, the perplexity increases don't lead to worse performance in the F1 metrics. On the contrary, a confidence score of 6, which translates to a GAP of 94.1%, performs the best in F1 and RF1 for the non-oracle model.

Model	**Confidence**	**PPL ↓**	**F1 ↑**	**RF1 ↑**	**GAP**
K2R BART	0	18.5	16.3	10.0	59.5
(LightWild+	2	19.1	16.4	10.2	78.4
LightQA KM)	6	20.2	16.4	10.3	94.1
	10	22.3	16.2	10.1	99.0
K2R BART	0	12.7	27.4	25.5	79.0
(oracle KM)	2	12.4	28.6	27.5	86.7
	6	12.1	29.9	29.2	94.2
	10	12.0	30.1	30.0	98.3

Table 20: Confidence-conditioned model on LightWild.

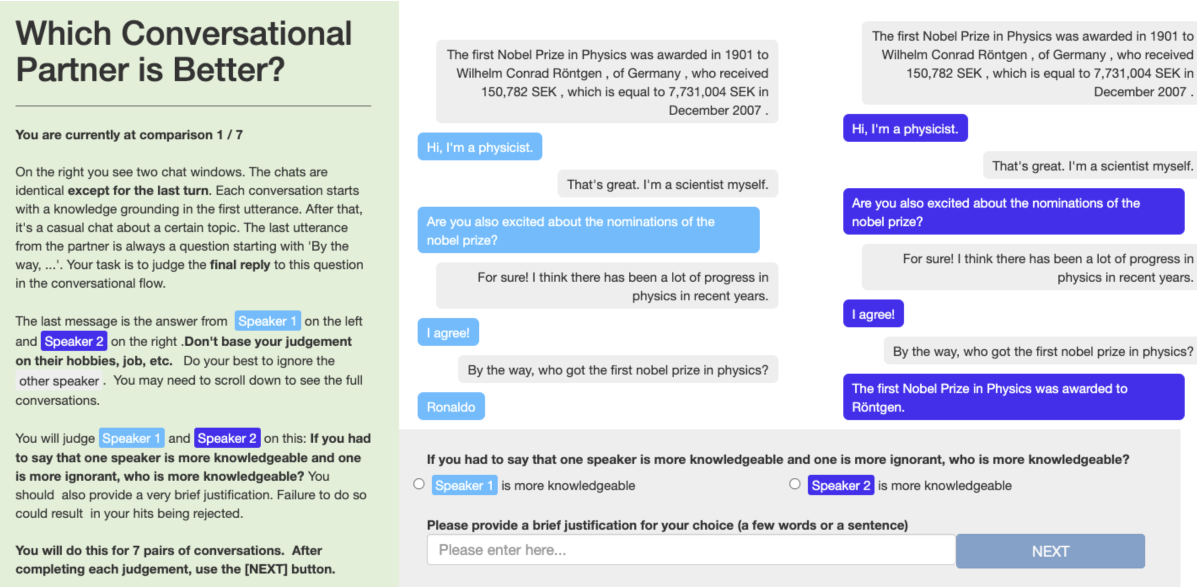
A.6 NQ Acute Eval Details

We closely follow the human evaluation setup studied by Li et al. (2019) and set up a pairwise model comparison on Amazon MTurk. To situate the NQ questions in a dialogue setting, we retrieve an episode from WoW where the chosen *topic* is mentioned in the question and use this as context. To have a smooth transition between dialogue context and the question itself, we prefix the question with "By the way, ...". The human evaluators are presented with a side-by-side comparison of the same context and question but with different answers corresponding to individual models. They are asked to read the dialogue and assess the final response according to one of the two following criteria, following the same wording as in (Li et al., 2019):

- If you had to say that one speaker is more knowledgeable and one is more ignorant, who is more knowledgeable?

- Who would you prefer to talk to for a long conversation?

In Figure 2 and 3, we provide screenshot examples of the interface used for the human evaluation. To ensure a high quality of evaluations, we only select people that manage to correctly solve two manually constructed onboarding examples.



Which Conversational Partner is Better?

You are currently at comparison 1 / 7

On the right you see two chat windows. The chats are identical **except for the last turn**. Each conversation starts with a knowledge grounding in the first utterance. After that, it's a casual chat about a certain topic. The last utterance from the partner is always a question starting with 'By the way, ...'. Your task is to judge the **final reply** to this question in the conversational flow.

The last message is the answer from Speaker 1 on the left and Speaker 2 on the right .**Don't base your judgement on their hobbies, job, etc.** Do your best to ignore the other speaker . You may need to scroll down to see the full conversations.

You will judge Speaker 1 and Speaker 2 on this: **If you had to say that one speaker is more knowledgeable and one is more ignorant, who is more knowledgeable?** You should also provide a very brief justification. Failure to do so could result in your hits being rejected.

You will do this for 7 pairs of conversations. After completing each judgement, use the [NEXT] button.

The first Nobel Prize in Physics was awarded in 1901 to Wilhelm Conrad Röntgen , of Germany , who received 150,782 SEK , which is equal to 7,731,004 SEK in December 2007 .

Hi, I'm a physicist.

That's great. I'm a scientist myself.

Are you also excited about the nominations of the nobel prize?

For sure! I think there has been a lot of progress in physics in recent years.

I agree!

By the way, who got the first nobel prize in physics?

Ronaldo

The first Nobel Prize in Physics was awarded in 1901 to Wilhelm Conrad Röntgen , of Germany , who received 150,782 SEK , which is equal to 7,731,004 SEK in December 2007 .

Hi, I'm a physicist.

That's great. I'm a scientist myself.

Are you also excited about the nominations of the nobel prize?

For sure! I think there has been a lot of progress in physics in recent years.

I agree!

By the way, who got the first nobel prize in physics?

The first Nobel Prize in Physics was awarded to Röntgen.

If you had to say that one speaker is more knowledgeable and one is more ignorant, who is more knowledgeable?

○ Speaker 1 is more knowledgeable ○ Speaker 2 is more knowledgeable

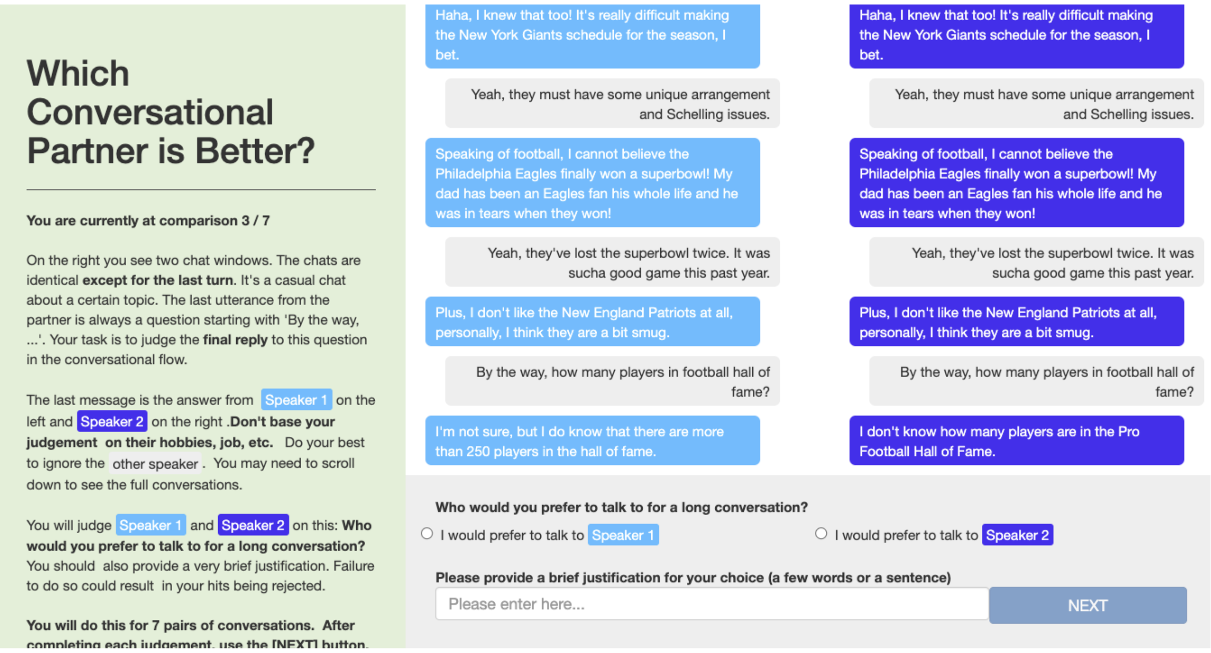
Please provide a brief justification for your choice (a few words or a sentence)

Please enter here...

Figure 2: Example interface for human evaluation for *knowledgeable*. The first utterance is a knowledge paragraph that answers the final question–provided to give the reviewer the relevant information to assess the models' answers. Then, there is a random dialogue roughly matching the topic of the final NQ question which is prefixed with "By the way, ...". The reviewer is asked to vote for the better response among the two models and provide a brief justification.

Challenger	Losses K2R	Wins K2R	K2R Win Reasons Sample	K2R Loss Reasons Sample
BART	18	91	Gives an answer with location. Precise and clear with proper response. Gives an answer with location. is more detailed The speaker gives a proper answer to the question.	Neither answers the question. They acknowledge what they don't know This speaker seems more correct gave the correct answer They have a lot more information stores
RAG DPR	26	86	Better Answer. He gives more in depth information More likely correct response. The level of detail is higher, and the phrasing is natural. The response actually answers the question.	gave a more up to date response knowledgeable but don't come off as a know it all They both were fine i just like 2s response better Neither answers the question. gave the correct answer
T5 (QA Model)	37	72	Both good, 2s response better though I prefer the longer reply Gives more detailed response. Give more information in their answer The level of detail is better.	The answer is more concise, and accurate. more direct answer This speaker answers the question directly The answer is more direct. more to the point

Table 21: Acute evaluation details for NQ on the question "If you had to say that one speaker is more knowledgeable and one is more ignorant, who is more knowledgeable?". The last two columns show some samples of justifications provided by human evaluators in the case of K2R winning and losing, respectively.



Which Conversational Partner is Better?

You are currently at comparison 3 / 7

On the right you see two chat windows. The chats are identical **except for the last turn**. It's a casual chat about a certain topic. The last utterance from the partner is always a question starting with 'By the way, ...'. Your task is to judge the **final reply** to this question in the conversational flow.

The last message is the answer from Speaker 1 on the left and Speaker 2 on the right .**Don't base your judgement on their hobbies, job, etc.** Do your best to ignore the other speaker . You may need to scroll down to see the full conversations.

You will judge Speaker 1 and Speaker 2 on this: **Who would you prefer to talk to for a long conversation?** You should also provide a very brief justification. Failure to do so could result in your hits being rejected.

You will do this for 7 pairs of conversations. After completing each judgement, use the [NEXT] button.

Haha, I knew that too! It's really difficult making the New York Giants schedule for the season, I bet.

Yeah, they must have some unique arrangement and Schelling issues.

Speaking of football, I cannot believe the Philadelphia Eagles finally won a superbowl! My dad has been an Eagles fan his whole life and he was in tears when they won!

Yeah, they've lost the superbowl twice. It was sucha good game this past year.

Plus, I don't like the New England Patriots at all, personally, I think they are a bit smug.

By the way, how many players in football hall of fame?

I'm not sure, but I do know that there are more than 250 players in the hall of fame.

Haha, I knew that too! It's really difficult making the New York Giants schedule for the season, I bet.

Yeah, they must have some unique arrangement and Schelling issues.

Speaking of football, I cannot believe the Philadelphia Eagles finally won a superbowl! My dad has been an Eagles fan his whole life and he was in tears when they won!

Yeah, they've lost the superbowl twice. It was sucha good game this past year.

Plus, I don't like the New England Patriots at all, personally, I think they are a bit smug.

By the way, how many players in football hall of fame?

I don't know how many players are in the Pro Football Hall of Fame.

Who would you prefer to talk to for a long conversation?

○ I would prefer to talk to Speaker 1 ○ I would prefer to talk to Speaker 2

Please provide a brief justification for your choice (a few words or a sentence)

Please enter here...

Figure 3: Example interface for human evaluation for *engaging*. We present the reviewer a random dialogue roughly matching the topic of the final NQ question which is prefixed with "By the way, ...". The reviewer is asked to vote for the better response among the two models and provide a brief justification.

A.7 Computational Setup

For all our experiments, we start from pre-trained checkpoints and only fine-tune the models to our specific tasks. To this end, we use up to four Tesla V100 Volta GPUs (32GB) in parallel for up to 24 hours for the Light experiments and up to 48 hours for the WoW and NQ experiments. For the hyperparameter search, we restrict ourselves to only search over the learning rate for up to four different values. In particular, we search over $[7 \times 10^{-6}, 5 \times 10^{-6}, 1 \times 10^{-5}, 3 \times 10^{-5}]$.

Challenger	Losses K2R	Wins K2R	K2R Win Reasons Sample	K2R Loss Reasons Sample
BART	30	93	It leads to a more thought-provoking conversation. The level of detail is higher, and the phrasing is natural. This person sounds more well-versed. the information is more worthwile stays on topic better	is less incorrect is confidently incorrect I prefer 1's phrasing acknowledges their uncertainty. sticks to the question more closely
RAG DPR	35	89	The answer is phrased better does a better job answering questions is more focused on responding to its partner sounds more well-versed in the conversation replies more naturally	seems more correct Provides a really insightful answer to the question more detailed in their explanations They have some similar preferences as me Neither answers the question.
T5 (QA Model)	41	83	I prefer complete sentence responses sounds better than simply giving the name More natural in the conversation The answer uses a full sentence. adds more to the conversation	is more concise The answer is more direct. the answer is less formal and fits the question better provides a more direct answer know the answer to the question

Table 22: Acute evaluation details for NQ on the question "Who would you prefer to talk to for a long conversation?". The last two columns show some samples of justifications provided by human evaluators in the case of K2R winning and losing, respectively.