

Controllable Factuality in Document-Grounded Dialog Systems Using a Noisy Channel Model

Nico Daheim^{1,*}, David Thulke^{2,3}, Christian Dugast³, Hermann Ney^{2,3}

¹Ubiquitous Knowledge Processing Lab

Department of Computer Science, Technical University of Darmstadt

²Chair for Human Language Technology and Pattern Recognition, RWTH Aachen University

³AppTek GmbH

www.ukp.tu-darmstadt.de

{thulke, ney}@i6.informatik.rwth-aachen.de

Abstract

In this work, we present a model for document-grounded response generation in dialog that is decomposed into two components according to Bayes' theorem. One component is a traditional ungrounded response generation model and the other component models the reconstruction of the grounding document based on the dialog context and generated response. We propose different approximate decoding schemes and evaluate our approach on multiple open-domain and task-oriented document-grounded dialog datasets. Our experiments show that the model is more factual in terms of automatic factuality metrics than the baseline model. Furthermore, we outline how introducing scaling factors between the components allows for controlling the tradeoff between factuality and fluency in the model output. Finally, we compare our approach to a recently proposed method to control factuality in grounded dialog, CTRL (Rashkin et al., 2021), and show that both approaches can be combined to achieve additional improvements.

1 Introduction

Recently, document-grounded dialog systems have seen an increase in popularity. Not only are they used to extend task-oriented systems beyond the narrow scope of fixed databases and APIs (Feng et al., 2020; Kim et al., 2020, 2021; Cohen et al., 2022), but also to ground open-domain conversations in information from the web (Zhou et al., 2018; Dinan et al., 2019; Komeili et al., 2022) or in persona descriptions to make dialog agents more interesting and engaging (Zhang et al., 2018). In any of these applications, the conversational system has to use the information from the document

and blend it into the response (Roller et al., 2020). This means that the response should remain fluent, grammatically correct and coherent to the previous utterances in the dialog on the one hand, and on the other hand that it correctly reflects the information from the document. This entails that no information is altered and also that no new information should be added if it is not immediately verifiable. This is especially crucial in all cases where a user uses a system to satisfy an information need (Santhanam et al., 2021).

While previous work has shown that retrieving relevant information is a crucial step for task-oriented (Kim et al., 2020) and open-domain document-grounded dialog systems and a potential mitigator of inconsistencies (Shuster et al., 2021), there is sufficient evidence that grounded response generation models may still fail to produce factual responses, even when the correct information is contained in its grounding document. In general, generating outputs that are both fluent and correct remains an open problem not only in dialog systems but natural language generation as a whole (Cao et al., 2018; Maynez et al., 2020; Roller et al., 2021; Ji et al., 2022) which potentially limits its industry adaptation. Models have been found to contradict themselves (Shuster et al., 2022b) or the grounding, and to add additional information that might be harmful and is not verifiable (Shuster et al., 2022a; Ji et al., 2022). Recently, different mitigation strategies were proposed. Cohen et al. (2022) for example use learned discriminators to decide from an n-best list while Rashkin et al. (2021) introduce special control tokens (Keskar et al., 2019) to encourage lexical overlap and entailment between grounding and response. However, discriminators based on estimating human judgments require additional data for training (Cohen et al., 2022) which is costly to obtain. Furthermore, discriminating based on single attributes holds the potential of harming other relevant properties. For

Code available at: https://github.com/ndaheim/noisy_channel_model

* Work done while at RWTH Aachen University, Chair for Human Language Technology and Pattern Recognition

example, discriminating based on factuality can lead to responses mostly repeating their grounding information (Cohen et al., 2022) and might introduce a loss of fluency and dialog coherence. In general, these two goals might be conflicting and their importance depends on the task at hand. While chit-chat lives from engagement and might tolerate inconsistencies, factuality is service-critical in task-oriented systems.

In this work, we present a probabilistic model that inherently combines both of these goals. By factorizing the model according to Bayes’ theorem, we obtain one component that models each goal explicitly. Hence, introducing scaling factors allows for controlling between them. Furthermore, additional unlabeled dialog data can be integrated easily to train one of its components. As directly decoding the model is intractable, we present different approximate decoding schemes for reranking and online decoding that yield significant gains in terms of automatic factuality metrics on several datasets.

2 Related Work

2.1 Document-grounded dialog systems

There has been significant work in document-grounded dialog systems in recent years. A large number of datasets have been proposed for open-domain dialog, in order to facilitate engaging conversations about a variety of topics, such as movies (Zhou et al., 2018), Wikipedia knowledge (Dinan et al., 2019; Dziri et al., 2022a), personal attributes of the agent (Zhang et al., 2018; Dinan et al., 2020) or arbitrary information from the internet (Komeili et al., 2022). Similarly, different task-oriented dialog datasets for information-seeking conversations have been proposed (Kim et al., 2020, 2021; Feng, 2021). Different works have dealt with the problem of document retrieval, for example on batching hard negatives (He et al., 2021) or efficient document retrieval (Thulke et al., 2021), as well as with identifying (Feng, 2021) or rephrasing salient passages within them (Shuster et al., 2022a). Finally, there is also significant work on generating grounded responses using this information, for example in low-resource scenarios (Zhao et al., 2020) or with an emphasis on faithful generations, which we will explore in the following section.

2.2 Hallucination in language generation and dialog

The problem of hallucinations, which one might define as information that is not grounded in the document, dialog context or by common sense, has recently received plenty of attention in neural language generation (Ji et al., 2022), for example in the field of summarization (Cao et al., 2018; Maynez et al., 2020) and dialog systems (Roller et al., 2021). Hence, different types of mitigation strategies that aim to increase the faithfulness of responses have been proposed. Notably, Gabriel et al. (2021) use a set of discriminators to rerank outputs from an n-best list for summarization. For dialog systems, Shuster et al. (2021) show that retrieving relevant information can reduce hallucinations. However, Santhanam et al. (2021) show that correct grounding does not guarantee faithful outputs. Cohen et al. (2022) train discriminators for dialog systems using human judgements, for example to encourage better grounded responses by adding high-quality responses found by the discriminator to the training data. Rashkin et al. (2021) augment the input of a grounded response generation model by additional control tokens (Keskar et al., 2019) to steer generations towards responses entailed by the grounding and Prabhumoye et al. (2021) add an additional attention mechanism to BART that focuses solely on the document.

Along with mitigation strategies, methods for model output and metric evaluation for factuality have been proposed. For example, Q^2 (Honovich et al., 2021) proposes a question-answering-based matching and BEGIN (Dziri et al., 2022c) a benchmark for metric evaluation.

Recently, Dziri et al. (2022b) also show that current grounded datasets contain ground-truth responses that further encourage hallucination by being insufficiently grounded and Dziri et al. (2022a) propose FaithDial as a filtered version of Wizard-of-Wikipedia (Dinan et al., 2019) that aims to mitigate this.

2.3 Noisy Channel Modeling in NLP

Given an input sequence x_1^T and output y_1^N , the noisy channel approach (Shannon, 1948) models the posterior probability of y_1^N given x_1^T as $p(y_1^N | x_1^T) = p(x_1^T | y_1^N) p(y_1^N) / p(x_1^T)$. For a long time, such models have been the dominant way of performing Automatic Speech Recognition (ASR) and Machine Translation (MT) (Brown et al.,

1993). In ASR, $p(x_1^T | y_1^N)$ models the acoustic channel (Bahl et al., 1983) and is often called channel model. With the advent of deep learning, discriminative approaches have become popular in both fields and achieve state-of-the-art results (Graves et al., 2006, 2013; Vaswani et al., 2017; Gulati et al., 2020). Nevertheless, the noisy channel approach has recently been explored again for MT (Yu et al., 2017; Yee et al., 2019; Yu et al., 2020; Jean and Cho, 2020; Subramanian et al., 2021), text classification (Min et al., 2022), style transfer (Thulke et al., 2022) and task-oriented dialog systems that are not document-grounded (Liu et al., 2021).

3 Grounded Response Generation

The goal of dialog systems is to find an appropriate system response u_{T+1} conditioned on a sequence of previous turns $u_1^T := (u_1, \dots, u_t, \dots, u_T)$ taken by different interlocutors, where each turn $u_t = [u_t]_0^{N_t} := ([u_t]_0, \dots, [u_t]_{N_t})$ is a sequence of N_t tokens from the model vocabulary \mathcal{V} prepended with the start of sequence symbol $[u_{T+1}]_0 := \langle \text{sos} \rangle$. This is usually done by means of a probabilistic language generation model that models the posterior distribution of the response given the context and is locally-normalized such that the response is generated autoregressively according to

$$p(u_{T+1} | u_1^T) = \prod_{n=1}^{N_{T+1}} p([u_{T+1}]_n | [u_{T+1}]_0^{n-1}, u_1^T). \quad (1)$$

In document-grounded dialog systems the response is additionally grounded in information that defines parts of its meaning and is given in the form of unstructured text. For example, in a restaurant booking setting, text on websites might show whether dogs can be brought or in a chit-chat system the grounding information might define the agent’s persona. While in the latter the grounding is known at test-time as a property of the agent, it is usually unknown in the former case. Then, the dialog system has to decide for relevant grounding documents, which are obtained from a document base \mathcal{D} . In this case, retrieval models that model the distribution $p(d | u_1^T, \mathcal{D})$ in order to rank documents $d \in \mathcal{D}$ may be employed. By introducing d as a latent variable, the posterior distribution of

the response given u_1^T and \mathcal{D} is given as follows.

$$\begin{aligned} p(u_{T+1} | u_1^T, \mathcal{D}) &= \sum_{d \in \mathcal{D}} p(u_{T+1}, d | u_1^T, \mathcal{D}) \\ &= \sum_{d \in \mathcal{D}} p(d | u_1^T, \mathcal{D}) \cdot p(u_{T+1} | u_1^T, d, \mathcal{D}) \\ &\approx \sum_{d \in \mathcal{D}} p(d | u_1^T, \mathcal{D}) \cdot p(u_{T+1} | u_1^T, d) \end{aligned}$$

For large \mathcal{D} the sum is then approximated by either top-k (Lewis et al., 2020b; Thulke et al., 2021) or maximum approximation. Furthermore, due to the input length restriction in current language generation models, the dependency on \mathcal{D} is usually dropped, as outlined in the last step. Using maximum approximation the model becomes

$$\begin{aligned} &\sum_{d \in \mathcal{D}} p(d | u_1^T, \mathcal{D}) \cdot p(u_{T+1} | u_1^T, d) \\ &\approx \max_{d \in \mathcal{D}} \{p(d | u_1^T, \mathcal{D}) \cdot p(u_{T+1} | u_1^T, d)\} \\ &\approx p(\hat{d} | u_1^T, \mathcal{D}) \cdot p(u_{T+1} | u_1^T, \hat{d}), \end{aligned}$$

where \hat{d} is the argument of the maximization of just $p(d | u_1^T, \mathcal{D})$ over \mathcal{D} since the simultaneous maximization over both components is intractable. This results in a widely-used two-step approach (e.g. (Kim et al., 2020; He et al., 2021)), where retrieval using $p(d | u_1^T, \mathcal{D})$ is followed by a response generation model that uses the retrieved document \hat{d} . Then, since the retrieval probability is constant during response generation, we may decide for a response according to the decision rule

$$(u_1^T, d) \mapsto \hat{u}_{T+1} = \arg \max_{u_{T+1} \in \mathcal{V}^*} \{p(u_{T+1} | u_1^T, d)\}. \quad (2)$$

In line with recent work (Yee et al., 2019; Liu et al., 2021), we will refer to the model used in Equation (2) as *direct model* in the following. While comparatively simple to use, we note that the model can not use data without grounding annotations directly and has been observed to produce a significant number of incorrect outputs, even with ground-truth grounding (Santhanam et al., 2021; Cohen et al., 2022).

4 A Noisy Channel Approach

We may use the following equivalent decision rule¹, where the direct model is factorized according to

¹A short proof can be found in Appendix A.1. Here, equivalence is meant with respect to the true distributions.

Bayes' Theorem:

$$\begin{aligned}
(u_1^T, d) &\mapsto \hat{u}_{T+1} = \arg \max_{u_{T+1} \in \mathcal{V}^*} \{p(u_{T+1} | u_1^T, d)\} \\
&= \arg \max_{u_{T+1} \in \mathcal{V}^*} \left\{ \underbrace{p(d | u_{T+1}, u_1^T)}_{\text{"channel model"}} \cdot \underbrace{p(u_{T+1} | u_1^T)}_{\text{"response generation model"}} \right\}.
\end{aligned} \tag{3}$$

Then, the first component can be formulated as a language generation model

$$\begin{aligned}
&p(d_1^M | u_{T+1}, u_1^T) \\
&= \prod_{m=1}^M p(d_m | d_0^{m-1}, u_{T+1}, u_1^T)
\end{aligned} \tag{4}$$

The first component can be interpreted as favouring responses that allow to reconstruct the grounding based on the generated response. Hence, one would expect accurate responses to receive higher probability estimates. The second component is an ungrounded response generation model as in Equation (1) and favours fluent responses irregardless of the grounding. Furthermore, it can be trained on large amounts of additional dialogues without grounding annotations. In line with previous work on similar models, we will refer to the first component as *channel model*. Introducing scaling factors between both components then allows for weighting the two objectives of correctness and fluency in order to control the outputs of the model.

We also note that one might arrive at a similar model by the following computation starting from the document-grounded dialog model without maximum approximation:

$$\begin{aligned}
&p(u_{T+1} | u_1^T, \mathcal{D}) \\
&= \sum_{d \in \mathcal{D}} p(d | u_1^T, \mathcal{D}) \cdot p(u_{T+1} | u_1^T, d, \mathcal{D}) \\
&= \sum_{d \in \mathcal{D}} p(d | u_1^T, \mathcal{D}) \cdot \frac{p(d | u_{T+1}, u_1^T, \mathcal{D}) \cdot p(u_{T+1} | u_1^T, \mathcal{D})}{p(d | u_1^T, \mathcal{D})} \\
&= \sum_{d \in \mathcal{D}} p(d | u_{T+1}, u_1^T, \mathcal{D}) \cdot p(u_{T+1} | u_1^T, \mathcal{D}).
\end{aligned} \tag{5}$$

Nevertheless, we leave the exploration of this model to future work.

Decoding the Noisy Channel Model Since the channel model has to be evaluated for each hypothesis generated by the response generation model, decoding the model even with beam search is intractable, as $k \cdot |\mathcal{V}|$ -many hypotheses would need

to be scored at each iteration with a beam size of k . Therefore, we derive two algorithms to approximately decode the model. First, we introduce *reranking*, where the noisy channel model is used to score a set of full candidate responses. Then, we introduce *online decoding*, where the noisy channel model is used to score partial responses during beam search. In both cases, we resort to a proposal model q to generate candidates. Recall, that we have the following relationship between direct and noisy channel model:

$$p(u_{T+1} | u_1^T, d) = \frac{p(d | u_{T+1}, u_1^T) \cdot p(u_{T+1} | u_1^T)}{p(d | u_1^T)}. \tag{6}$$

During decoding with the maximum-approximated model introduced in the previous section, $p(d | u_1^T)$ is constant and may be dropped. This makes the direct model the natural choice for a proposal model as we would get the same results as with the noisy channel model given the true distributions.

In reranking, the direct model then generates a set of full responses U_{T+1} , from which we decide according to the noisy channel model as follows:

$$\hat{u}_{T+1} = \arg \max_{u_{T+1} \in U_{T+1}} \left\{ p(d | u_{T+1}, u_1^T)^{\lambda_1} \cdot p(u_{T+1} | u_1^T)^{\lambda_2} \right\}, \tag{7}$$

where $\lambda_1 \in \mathbb{R}_{\geq 0}$ and $\lambda_2 \in \mathbb{R}_{\geq 0}$ are scaling factors. One might also add the direct model without additional computational effort (the probabilities are already calculated during beam search) as an additional factor which has shown beneficial in earlier works (Liu et al., 2021). We note that this resembles the use of discriminators to select responses but does not require additional annotation effort beyond the grounding annotations.

Since reranking is limited by the hypotheses generated by the proposal model, we propose an online decoding algorithm in which the noisy channel model is used during beam search. Since the channel model $p(d | u_{T+1}, u_1^T)$ depends on the final hypothesis that is not available during search, we approximate it using a model $p(d | [u_{T+1}]_0^n, u_1^T)$ that only depends on partially generated hypotheses similar to Liu et al. (2021). At each step, we score the k hypotheses in our beam using the noisy channel model. Since scoring all $k \cdot |V|$ possible extensions is infeasible, we select the best k extensions only using their direct model score. The noisy channel model score for each of the k partial hypotheses $[u_{T+1}]_0^n$ up to length n is then calculated

Algorithm 1 Pseudocode for Liu et al.’s (-) (Liu et al., 2021) and our (+) decoding algorithm.

Input Grounding d , dialog context u_1^T , beam sizes k_1, k_2

Output Response u_{T+1}

```

Beam:  $\mathcal{B} = \{\langle \text{sos} \rangle\}$ 
score( $w$ ) =  $\log p(w \mid u_1^T, d)$ 
           +  $\lambda_1 \cdot \log p(d \mid w, u_1^T)$ 
           +  $\lambda_2 \cdot \log p(w \mid u_1^T)$ 

 $q(v, w) = \log p(v \mid w, u_1^T, d)$ 

while end( $\mathcal{B}$ ) is False do
-    $\mathcal{B}' = \emptyset$ 
-   for  $w \in \mathcal{B}$  do
-      $\mathcal{B}' = \mathcal{B}' \cup \{w \circ v \mid v \in \text{top-}k_1\{q(v, w)\}\}$ 
-   end for
-    $\mathcal{B} = \text{top-}k_2\{\text{score}(w)\}$ 
+    $\mathcal{B} = \text{top-}k_2\{q(v, w) + \text{score}(w)\}$ 
       $\begin{matrix} w \in \mathcal{B}' \\ w \circ v: \\ w \in \mathcal{B}, v \in \mathcal{V} \end{matrix}$ 
end while

 $\hat{u}_{T+1} = \arg \max_{u_{T+1} \in \mathcal{B}} \left\{ \frac{\text{score}(u_{T+1})}{\text{len}(u_{T+1})} \right\}$ 

```

as follows:

$$p(d \mid [u_{T+1}]_0^n, u_1^T)^{\lambda_1} \cdot p([u_{T+1}]_0^n \mid u_1^T)^{\lambda_2}. \quad (8)$$

Again, we might add the direct model as an additional factor to the score. The algorithm requires locally-normalized models and a channel model trained on partial responses, which we obtain by truncating responses according to a uniform distribution over their length in training.

Lastly, we experiment with the algorithm proposed by Liu et al. (2021) which uses the direct model to generate k_1 extensions to each of the k_2 hypotheses in the beam. The $k_1 \cdot k_2$ new hypotheses are then pruned back to size k_2 using the noisy channel model. We outline both algorithms in Algorithm 1.

5 Experiments

We evaluate our model on multiple different English document-grounded dialog datasets by comparing it to the direct modeling objective and the CTRL model (Keskar et al., 2019) presented by Rashkin et al. (2021), where the input is augmented by a sequence of control tokens c_1^m . Thus, the model becomes $p(u_{T+1} \mid u_1^T, d, c_1^m)$ and we use the noisy channel model $p(d \mid u_{T+1}, u_1^T, c_1^m) \cdot p(u_{T+1} \mid u_1^T)$ in line with Section 4. We omit the "objective voice" token, since, for example, Personachat specifically targets conversations where the system responds in first person. The datasets and metrics are described in Section 5.1 and 5.2,

respectively. In all our experiments, we finetune the BART-large (Lewis et al., 2020a) checkpoint that is provided as part of the huggingface transformers (Wolf et al., 2020) library, which we further use to implement our experiments. In order to determine the scaling factors for the Noisy Channel model, we do a hyperparameter sweep across $\lambda_i \in \{0.1, 0.2, \dots, 2.0\}$ on the validation sets and choose the parameters that perform best in terms of Q^2 . We use $\lambda_1 = 0.6, \lambda_2 = 0.4$ for online decoding and $\lambda_1 = 0.5, \lambda_2 = 0.2$ for reranking for all experiments after seeing similar trends on all datasets. Furthermore, we always use our proposed online decoding algorithm (see Section 6.3 for a comparison). The results obtained with these experiments are discussed in Section 6.

5.1 Datasets

This section gives a brief overview of the different datasets used in our experiments which capture a variety of settings, for example task-oriented and open-domain dialogs grounded in documents of varying lengths. Dataset statistics can be found in Appendix A.3.

Personachat Personachat (Zhang et al., 2018) is a crowdsourced open-domain dialog dataset, where dialogs are grounded in persona descriptions that consist of five short sentences. In our experiments, we use the *self* configuration and evaluate on the validation split.

Wizard-of-Wikipedia Wizard-of-Wikipedia (WoW) (Dinan et al., 2019) is a crowdsourced open-domain dialog dataset, where turns are grounded in Wikipedia articles. Only the wizard can access the grounding in order to teach the apprentice but also choose not to use any grounding. We evaluate on the subset of grounded wizard turns.

DSTC9 DSTC9 Track 1 is an extension of the MultiWoz 2.1 dataset (Eric et al., 2020), where turns require information beyond the existing API structure and which was collected from FAQ documents. The test set contains conversations about a new location and a held-out domain, as well as transcripts of spoken conversations.

Doc2Dial Doc2dial (Feng et al., 2020) is a task-oriented dialog dataset, where the agent provides a user with information from public government service websites. The grounding annotations are

Method	Personachat					FaithDial				
	sBLEU (u_{T+1}, \hat{u}_{T+1})	METEOR	Q^2	BERTScore (u_{T+1}, d)	F1	sBLEU (u_{T+1}, \hat{u}_{T+1})	METEOR	Q^2	BERTScore (u_{T+1}, d)	F1
Direct Model	5.59	17.2	0.35	0.083	14.0	15.16	41.0	0.86	0.605	69.9
+ Reranking	5.46	17.3	0.48	0.162	22.9	14.36	40.1	0.89	0.654	74.4
+ Onl. Decoding	5.60	19.1	0.47	0.184	25.6	13.73	39.3	0.89	0.685	77.6
CTRL	4.87	15.6	0.63	0.256	30.6	13.65	38.0	0.92	0.725	77.8
+ Reranking	4.14	14.4	0.70	0.303	35.7	13.19	37.9	0.93	0.749	80.6
+ Onl. Decoding	4.65	16.5	0.65	0.320	40.1	12.70	37.1	0.92	0.749	81.4
	WoW test_unseen					WoW test_seen				
Direct Model	18.68	37.7	0.87	0.661	74.3	18.32	36.9	0.86	0.670	74.5
+ Reranking	18.82	37.6	0.90	0.707	78.5	18.47	36.8	0.89	0.702	75.3
+ Onl. Decoding	17.98	37.1	0.89	0.709	79.6	17.98	37.1	0.89	0.720	79.8
CTRL	17.95	35.8	0.93	0.715	77.2	17.49	35.0	0.92	0.729	78.0
+ Reranking	17.89	35.8	0.93	0.740	79.8	17.67	35.0	0.92	0.746	80.1
+ Onl. Decoding	17.14	35.3	0.92	0.743	81.3	17.38	35.3	0.92	0.766	82.6
	DSTC9					Doc2Dial				
Direct Model	18.60	41.5	0.71	0.439	61.9	50.08	63.0	0.78	0.680	80.4
+ Reranking	18.33	41.3	0.76	0.489	68.4	49.90	62.3	0.83	0.725	84.3
+ Onl. Decoding	17.72	42.4	0.75	0.473	68.8	49.34	62.9	0.80	0.713	83.3
CTRL	18.63	42.6	0.76	0.481	70.3	50.30	61.8	0.85	0.754	85.9
+ Reranking	18.45	41.6	0.81	0.531	75.2	50.31	61.4	0.89	0.796	88.7
+ Onl. Decoding	17.92	42.8	0.79	0.506	74.3	49.11	61.4	0.87	0.764	87.5

Table 1: Main results of our model compared to the direct model and CTRL (Rashkin et al., 2021). We use our online decoding algorithm and all results are within an effective beam size of 30.

given on different levels. We use a concatenation of the annotated spans as a grounding instead of taking entire paragraphs.

FaithDial Based on the observation that current dialog datasets contain insufficiently grounded annotations that encourage hallucinations (Dziri et al., 2022b), Dziri et al. (2022a) release FaithDial, an edited version of Wizard-of-Wikipedia that contains significantly less hallucinations. For this, crowdworkers have edited 44% of the grounded training responses from seeker-initiated conversations and all those from the validation and test set.

5.2 Evaluation Metrics

In line with the shared tasks on some of the datasets we use for evaluation (Kim et al., 2020, 2021; Feng et al., 2020), we use the sacrebleu (sBLEU) (Post, 2018) implementation of BLEU (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005) to assess our model generations with word-overlap based metrics. In addition to that, we use BERTScore (Zhang et al., 2020) and the token-level F1-Score between u_{T+1} and d , as well as the recently proposed Q^2 metric (Honovich et al., 2021) to evaluate the factual consistency of our models. Q^2 is a model-based metric that matches the answers, which are derived from response and grounding, to questions generated based on the response using an NLI model and has shown strong correlations with human judgements in system-level

evaluation on WoW.

5.3 Retrieval

In addition to experiments that use ground-truth grounding, we also experiment with using the outputs of retrieval models, since usually the grounding is not known at inference time. For retrieval we use two architectures. First, a Bi-Encoder (Bromley et al., 1993), where a dialog and document encoder model map u_1^T and each $d \in \mathcal{D}$ to a fixed-size dense vector of the same dimension, respectively. The grounding document is determined by nearest neighbor search, i.e. the decision is made for the document whose vector is closest to the dialog vector in the embedding space. In our experiments, the weights of dialog and document encoder are shared and trained using the Triplet loss criterion.

Secondly, we employ a Cross-Encoder which provides strong performance across a variety of tasks but remains too inefficient in order to be used in practice with large \mathcal{D} (Reimers and Gurevych, 2019; Humeau et al., 2020; Karpukhin et al., 2020; Thulke et al., 2021). In the Cross-Encoder, dialog context and document are concatenated as the input to a Transformer model that subsequently performs relevance classification such that the document with the highest score is retrieved.

We use RoBERTa-large (Liu et al., 2019) for all experiments use Recall@1 (R@1) for evaluation.

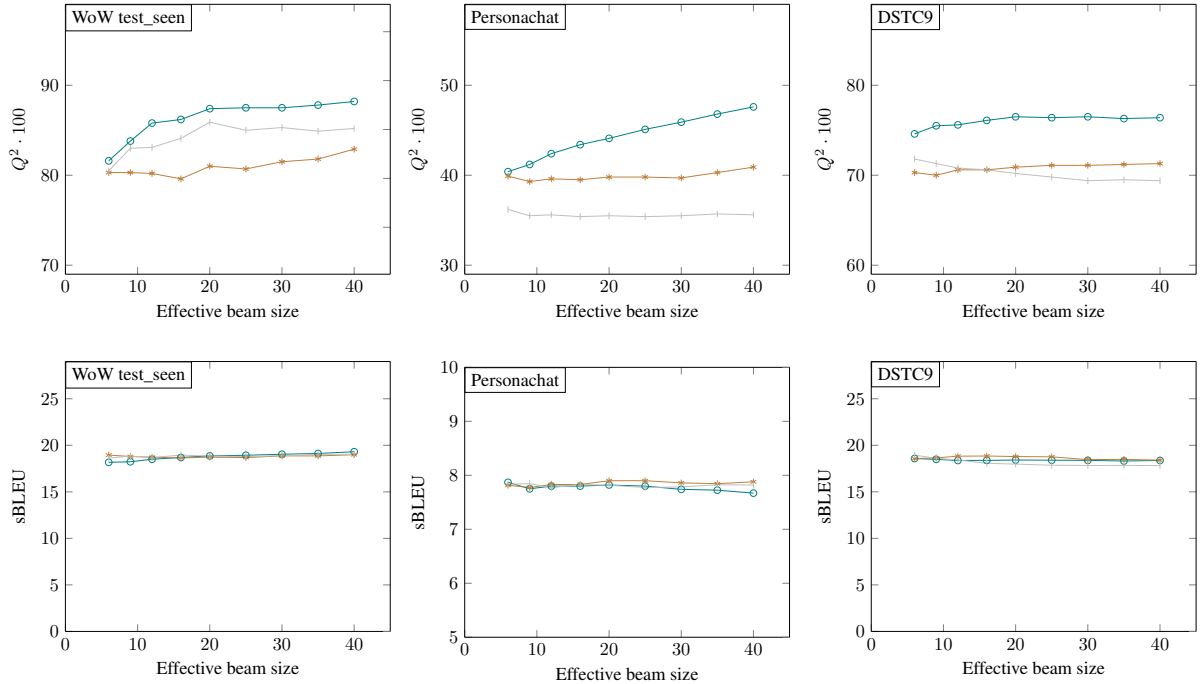


Figure 1: Q^2 and sBLEU for the noisy channel model using our decoding (o), the one proposed by Liu et al. 2019 (*), where $k_1 \cdot k_2$ is the effective beam size, and direct model beam search (l) by compute budget.

6 Results

Table 1 shows the results obtained with our model using our proposed online decoding algorithm, reranking, and no additional data in training but the corresponding training set. We identify the following trends:

1) Our model consistently outperforms the direct modeling objective in terms of all automated factuality metrics.

2) There is no clear trend in terms of word-overlap-based metrics, where our model and the direct model show comparable performance.

3) CTRL gives larger improvements in terms of Q^2 on all datasets but DSTC9. Nevertheless, the additional control tokens may be seen as a data filtering method that adds new information to the training data that is not available to our model.

4) Adding control tokens to our model, i.e. combining CTRL with the noisy channel approach, gives further improvements in terms of factuality metrics.

5) The improvements obtained on the unseen set of Wizard-of-Wikipedia indicate that the model is also able to generalize appropriately to new information.

Overall, the results indicate that by scaling the channel model contribution appropriately, the faith-

λ_1/λ_2	sBLEU \uparrow (u_{T+1}, \hat{u}_{T+1})	PPL \downarrow	Q^2 \uparrow	BERTScore \uparrow (u_{T+1}, d)	LCS
Direct Model	5.59	9.6	0.36	0.08	0.33
0.0 / 0.5	5.26	9.1	0.34	0.07	0.33
0.2 / 0.5	5.67	10.2	0.38	0.12	0.35
0.4 / 0.5	5.52	10.7	0.40	0.13	0.36
0.6 / 0.5	5.48	11.2	0.41	0.14	0.37
0.8 / 0.5	5.29	11.5	0.42	0.15	0.38
1.0 / 0.5	5.07	11.9	0.43	0.16	0.38
1.0 / 5.0	3.94	8.70	0.38	0.17	0.33
CTRL	4.89	17.2	0.64	0.27	0.66
0.6 / 0.4	4.70	19.5	0.64	0.32	0.62
Gold response	100.0	18.3	0.26	0.09	0.25
Gold document	1.97	47.8	99.7	1.0	1.0

Table 2: Q^2 and Perplexity by ratio of factors on Personachat with beam size 10.

fulness of responses can indeed be improved in comparison to the direct model. In the following, we present further results to understand how the model behaves under different scaling factors (Section 6.1), uncertain retrieval (Section 6.2), different compute budgets (Section 6.3), and the presence of additional data (Section 6.4) before concluding the section with a qualitative analysis (Section 6.5).

6.1 Controllability

Table 2 shows results obtained with noisy channel online decoding with different scaling factors. In addition to the previously mentioned metrics, we report Perplexity as a proxy for fluency (Dinan et al., 2020) and the longest common subsequence

Method	R@1	sBLEU	METEOR	Q^2	BERTScore	F1
Bi-Encoder	83.0	-	-	-	-	-
Direct Model	83.0	17.87	39.8	0.66	0.403	56.2
Online Decoding	83.0	17.80	41.0	0.68	0.434	61.3
Cross-Encoder	87.2	-	-	-	-	-
Direct Model	87.2	18.04	40.4	0.66	0.412	57.7
Online Decoding	87.2	17.83	41.6	0.69	0.446	63.1

Table 3: Results on the outputs of a Bi-Encoder and Cross-Encoder retrieval model on DSTC9 test.

(LCS) of u_{T+1} and d normalized by the length of u_{T+1} to indicate how much of the response is copied from the grounding. We can see that with a higher channel model factor, the model copies more from the grounding and produces more factual responses. However, these responses have a higher perplexity, as judged by a response generation model trained on all datasets. Thus, by differing the scaling factors one might control how much from the grounding is copied and how much importance is given to retaining a response that fits well into the dialog. CTRL copies significantly more from the grounding than our model. All models, including the direct model, copy more than the ground-truth responses, which is also consistent across all datasets. We note that this may not be desirable in open-domain dialog settings.

6.2 Retrieval

Table 3 show results obtained with the noisy channel model and the direct model with uncertain retrieval. We limit our study to DSTC9, since setting-up competitive retrievers for each dataset is non-trivial. For example, on doc2dial a QA-based pipeline is often used (Daheim et al., 2021; Feng, 2021)). For both the Bi-Encoder and the better performing Cross-Encoder retrieval model, the noisy channel model produces more factual responses than the direct model. This shows that the model remains effective in a realistic setting.

6.3 Online Decoding by compute budget

Figure 1 shows a comparison of our proposed decoding algorithm, the algorithm outlined in (Liu et al., 2021) and standard beam search using the direct model. We can see that the noisy channel model with our decoding algorithm consistently gives the best performance in terms of Q^2 . Furthermore, the direct model even performs better than the decoding algorithm proposed in (Liu et al., 2021) on Wizard-of-Wikipedia for the same compute budget, which takes a significant number of beams for improvements to show. In general, a

Method	Personachat		DSTC9		FaithDial	
	sBLEU	Q^2	sBLEU	Q^2	sBLEU	Q^2
Reranking	5.46	0.48	18.33	0.76	14.36	0.89
+ data	5.89	0.46	18.47	0.73	15.29	0.87
Onl. Decoding	5.60	0.47	17.72	0.75	13.73	0.89
+ data	5.60	0.46	18.67	0.77	13.51	0.91

Table 4: Results with noisy channel model and additional response generation model training data.

larger beam size improves the performance of the noisy channel model but not the direct model.

6.4 Additional data

In order to study the effect of additional training data for the response generation model, we train the component on all target datasets as well as MultiWoZ 2.1 (Eric et al., 2020)², Taskmaster-1 (Byrne et al., 2019), TopicalChat (Gopalakrishnan et al., 2019) and CMU DoG (Zhou et al., 2018). However, as shown in Table 4, we do not see consistent but often dataset-specific improvements. For example, Q^2 tends to be better with more data in reranking but not online decoding, which also did not change in our experiments with different scaling factors.

6.5 Qualitative Analysis

When comparing the generation outputs of all models, we can make the following observations:

1) the outputs of both the noisy channel model and CTRL appear more faithful to the grounding and more specific. For example, on DSTC9 and doc2dial the direct model sometimes leaves out important details, such as that an ID card needs to be shown when someone picks up a ticket at a train station, which both noisy channel model and CTRL incorporate into the response. Furthermore, the direct model appears to generate generic responses more often, such as "do you have a pet?" when the topic in Personachat is "dog" or "cat".

2) In general, qualitative analysis supports the results from our automatic evaluation that a higher channel model factor implies more from the grounding being copied into the response, which however can come at a loss of fluency and coherence.

3) A higher response generation model factor leads to more abstractiveness and a better connected response that, for example, contains follow-up questions more often. On the other hand, a too high factor also led to hallucinations in our experiments.

²We remove the dialogs contained in DSTC9.

4) When comparing CTRL and the noisy channel model, the main difference appears to be that the responses of the noisy channel model (with a suitable factor) seem more connected to the previous turns than in CTRL, especially on a chit-chat task like Personachat. Nevertheless, with a higher channel model factor the generations become more similar to those of CTRL. The combination of CTRL and the noisy channel model might be especially suited for task-oriented dialog, where faithfulness is crucial, whereas for some open-domain settings we think that our model may be more suitable without control tokens.

Finally, some example outputs can be found in the Appendix.

7 Conclusion

In this paper, we present a model for response generation in document-grounded dialog that explicitly optimizes for faithfulness and fluency. The model decomposes the posterior distribution of response given context and grounding into two components according to Bayes' Theorem and thus, by introducing scaling factors, allows for encouraging more correct or more fluent responses. Since decoding the model directly is intractable, we derive and compare different approximate decoding schemes that use reranking or online decoding. We compare the model to directly modeling the posterior distribution of response given context and grounding and a variant of CTRL that was proposed recently to encourage faithfulness in grounded response generation, which we again factorize into two components. An evaluation on five different open-domain and task-oriented dialog datasets shows improvements in terms of factuality on top of both models. Furthermore, we highlight how the scaling factors can be used to control how much the model copies from the grounding and how much weight is given to a well-connected response, and we investigate the influence of additional training data for one of the model components. Lastly, we show that our model also gives improvements when dealing with uncertain document retrieval.

7.1 Future Work

In future work we would like to explore the noisy channel model for document-grounded response generation without maximum approximation, such as the one that we present in Equation (5).

8 Limitations

The main limitation of our model is that the approximate decoding schemes introduce significant computational overhead in comparison to the direct model and CTRL, which amounts to a factor of up to 10 for online decoding with a large beam size using a not yet fully optimized implementation. Furthermore, decoding is a lot more complex and the scaling factors need to be tuned which implies significant additional computations that are necessary and results in a larger carbon footprint. In addition, our noisy channel model has a significantly higher number of parameters than the baseline model. Finally, we mainly rely on automatic metrics to assess the faithfulness of the proposed approaches and leave out a broader evaluation of the general quality of generated responses but for a small-scale qualitative study cf. Section 6.5.

9 Broader Impact

In general, generative dialog systems are a promising field of research and can be less restricted in the topics they can deal with than rule-based models or approaches that use predefined dialog flows, for example, which require a lot of hand-crafting and possibly experts to write appropriate responses. Nevertheless, such language generation-based approaches bear the danger of repeating harmful content and biases that may have been present in the training data or of generating inappropriate responses, in general. Furthermore, some applications require faithful responses by law or are at least critical to the service. While our model can improve the faithfulness of responses, it can not be guaranteed. Therefore, we would not recommend to use the model in these applications. While document-grounded dialog systems in general can be used to ground system responses in helpful and correct real-world information, these systems could potentially also be misused to ground dialog systems in misinformation or other harmful documents.

Acknowledgements

This work has been financially supported by multiple sources: The project HYKIST funded by the German Federal Ministry of Health on the basis of a decision of the German Federal Parliament (Bundestag) under funding ID ZMVI1-2520DAT04A.

The project NeuroSys which, as part of the initiative "Clusters4Future", is funded by the Fed-

eral Ministry of Education and Research BMBF (03ZU1106DA).

The German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

The European Union under the Horizon Europe grant No. 101070351 (SERMAS).

References

- Lalit R. Bahl, Frederick Jelinek, and Robert L. Mercer. 1983. [A Maximum Likelihood Approach to Continuous Speech Recognition](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2):179–190.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. [Signature Verification Using a Siamese Time Delay Neural Network](#). In *Advances in Neural Information Processing Systems 6, [7th NIPS Conference, Denver, Colorado, USA, 1993]*, pages 737–744. Morgan Kaufmann.
- Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. [The Mathematics of Statistical Machine Translation: Parameter Estimation](#). *Computational Linguistics*, 19(2):263–311.
- Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. [Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset](#). 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 4516–4525, Hong Kong, China. Association for Computational Linguistics.
- Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. [Faithful to the Original: Fact Aware Neural Abstractive Summarization](#). 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 4784–4791. AAAI Press.
- Aaron Daniel Cohen, Adam Roberts, Alejandra Molina, Alena Butryna, Alicia Jin, Apoorv Kulshreshtha, Ben Hutchinson, Ben Zevenbergen, Blaise Hilary Aguera-Arcas, Chung ching Chang, Claire Cui, Cosmo Du, Daniel De Freitas Adiwardana, Dehao Chen, Dmitry (Dima) Lepikhin, Ed H. Chi, Erin Hoffman-John, Heng-Tze Cheng, Hongrae Lee, Igor Krivokon, James Qin, Jamie Hall, Joe Fenton, Johnny Soraker, Kathy Meier-Hellstern, Kristen Olson, Lora Moïs Aroyo, Maarten Paul Bosma, Marc Joseph Pickett, Marcelo Amorim Menegali, Marian Croak, Mark Díaz, Matthew Lamm, Maxim Krikun, Meredith Ringel Morris, Noam Shazeer, Quoc V. Le, Rachel Bernstein, Ravi Rajakumar, Ray Kurzweil, Romal Thoppilan, Steven Zheng, Taylor Bos, Toju Duke, Tulsee Doshi, Vinodkumar Prabhakaran, Will Rusch, YaGuang Li, Yanping Huang, Yanqi Zhou, Yuanzhong Xu, and Zhifeng Chen. 2022. [LaMDA: Language Models for Dialog Applications](#). *ArXiv preprint*, abs/2201.08239.
- Nico Daheim, David Thulke, Christian Dugast, and Hermann Ney. 2021. [Cascaded span extraction and response generation for document-grounded dialog](#). In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, pages 57–62, Online. Association for Computational Linguistics.
- Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS '18 Competition*, pages 187–208, Cham. Springer International Publishing.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Nouha Dziri, Ehsan Kamaloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo Ponti, and Siva Reddy. 2022a. [FaithDial: A Faithful Benchmark for Information-Seeking Dialogue](#). *ArXiv preprint*, abs/2204.10757.
- Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022b. [On the origin of hallucinations in conversational models: Is it the datasets or the models?](#) In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.
- Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022c. [Evaluating Attribution in Dialogue Systems: The BEGIN Benchmark](#). *Transactions of the Association for Computational Linguistics*, 10:1066–1083.

- Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. [MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines](#). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.
- Song Feng. 2021. [DialDoc 2021 Shared Task: Goal-Oriented Document-grounded Dialogue Modeling](#). In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, pages 1–7, Online. Association for Computational Linguistics.
- Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. [doc2dial: A Goal-Oriented Document-Grounded Dialogue Dataset](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. Association for Computational Linguistics.
- Saadia Gabriel, Antoine Bosselut, Jeff Da, Ari Holtzman, Jan Buys, Kyle Lo, Asli Celikyilmaz, and Yejin Choi. 2021. [Discourse Understanding and Factual Consistency in Abstractive Summarization](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 435–447, Online. Association for Computational Linguistics.
- Karthik Gopalakrishnan, Behnam Hedayatnia, Qiang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. [Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations](#). In *Proc. Interspeech 2019*, pages 1891–1895.
- Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. 2006. [Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks](#). In *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, volume 148 of *ACM International Conference Proceeding Series*, pages 369–376. ACM.
- Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. [Speech recognition with deep recurrent neural networks](#). In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649.
- Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. [Conformer: Convolution-augmented Transformer for Speech Recognition](#). In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 5036–5040. ISCA.
- Huang He, Hua Lu, Siqi Bao, Fan Wang, Hua Wu, Zhengyu Niu, and Haifeng Wang. 2021. [Learning to Select External Knowledge with Multi-Scale Negative Sampling](#). *ArXiv preprint*, abs/2102.02096.
- Or Honovich, Leshem Choshen, Roei Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. [Q²: Evaluating Factual Consistency in Knowledge-Grounded Dialogues via Question Generation and Question Answering](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- 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 *International Conference on Learning Representations*.
- Sébastien Jean and Kyunghyun Cho. 2020. [Log-Linear Reformulation of the Noisy Channel Model for Document-Level Neural Machine Translation](#). In *Proceedings of the Fourth Workshop on Structured Prediction for NLP*, pages 95–101, Online. Association for Computational Linguistics.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. [Survey of Hallucination in Natural Language Generation](#). *ArXiv preprint*, abs/2202.03629.
- 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.
- Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. [CTRL - A Conditional Transformer Language Model for Controllable Generation](#). *ArXiv preprint*, abs/1909.05858.
- Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur. 2020. [Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access](#). In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 278–289, 1st virtual meeting. Association for Computational Linguistics.
- Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papanagelis, Karthik Gopalakrishnan, Behnam Hedayatnia, and Dilek Hakkani-Tur. 2021. [“how robust r u?”: Evaluating task-oriented dialogue systems on spoken conversations](#). In *ASRU 2021*.

- Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. [Internet-augmented dialogue generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland. 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 Pretraining 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*.
- Qi Liu, Lei Yu, Laura Rimell, and Phil Blunsom. 2021. [Pretraining the Noisy Channel Model for Task-Oriented Dialogue](#). *Transactions of the Association for Computational Linguistics*, 9:657–674.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A Robustly Optimized BERT Pretraining Approach](#). *ArXiv preprint*, abs/1907.11692.
- 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.
- Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. [Noisy channel language model prompting for few-shot text classification](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5316–5330, Dublin, Ireland. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a Method for Automatic Evaluation of Machine Translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Matt Post. 2018. [A Call for Clarity in Reporting BLEU Scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Shrimai Prabhumoye, Kazuma Hashimoto, Yingbo Zhou, Alan W Black, and Ruslan Salakhutdinov. 2021. [Focused Attention Improves Document-Grounded Generation](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4274–4287, Online. Association for Computational Linguistics.
- Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. [Increasing Faithfulness in Knowledge-Grounded Dialogue with Controllable Features](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 704–718, Online. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Stephen Roller, Y-Lan Boureau, Jason Weston, Antoine Bordes, Emily Dinan, Angela Fan, David Gunning, Da Ju, Margaret Li, Spencer Poff, Pratik Ringshia, Kurt Shuster, Eric Michael Smith, Arthur Szlam, Jack Urbanek, and Mary Williamson. 2020. [Open-Domain Conversational Agents: Current Progress, Open Problems, and Future Directions](#).
- 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.
- Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tur. 2021. [Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation](#). In *EMNLP 2021 Workshop on NLP for Conversational AI*.
- C. E. Shannon. 1948. [A mathematical theory of communication](#). *The Bell System Technical Journal*, 27(3):379–423.
- Kurt Shuster, Mojtaba Komeili, Leonard Adolphs, Stephen Roller, Arthur Szlam, and Jason Weston. 2022a. [Language Models that Seek for Knowledge: Modular Search & Generation for Dialogue and Prompt Completion](#).

- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. [Retrieval Augmentation Reduces Hallucination in Conversation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Kurt Shuster, Jack Urbanek, Arthur Szlam, and Jason Weston. 2022b. [Am I me or you? state-of-the-art dialogue models cannot maintain an identity](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2367–2387, Seattle, United States. Association for Computational Linguistics.
- Sandeep Subramanian, Oleksii Hrinchuk, Virginia Adams, and Oleksii Kuchaiev. 2021. [NVIDIA NeMo’s neural machine translation systems for English-German and English-Russian news and biomedical tasks at WMT21](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 197–204, Online. 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](#). In *Workshop on DSTC9, AAAI*.
- David Thulke, Nico Daheim, Christian Dugast, and Hermann Ney. 2022. [Adapting document-grounded dialog systems to spoken conversations using data augmentation and a noisy channel model](#). In *Workshop on DSTC10, AAAI*.
- 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.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-Art Natural Language Processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Kyra Yee, Yann Dauphin, and Michael Auli. 2019. [Simple and Effective Noisy Channel Modeling for Neural Machine Translation](#). 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 5696–5701, Hong Kong, China. Association for Computational Linguistics.
- Lei Yu, Phil Blunsom, Chris Dyer, Edward Grefenstette, and Tomas Kocisky. 2017. [The Neural Noisy Channel](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer. 2020. [Better Document-Level Machine Translation with Bayes’ Rule](#). *Transactions of the Association for Computational Linguistics*, 8:346–360.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. [Personalizing Dialogue Agents: I have a dog, do you have pets too?](#) In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating Text Generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020. [Low-Resource Knowledge-Grounded Dialogue Generation](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- 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 Appendix

A.1 Proof of Equivalence of direct and noisy channel model decoding

Given a dialog context u_1^T and grounding d , we may define the search problem as finding an appropriate response u_{T+1} . Following our derivation from Section 3 we can formalize this according to the decision rule

$$(u_1^T, d) \mapsto \hat{u}_{T+1} = \arg \max_{u_{T+1} \in \mathcal{V}^*} \{p(u_{T+1} | u_1^T, d)\}, \quad (9)$$

where \mathcal{V}^* denotes the set of finite strings that may be constructed from a fixed vocabulary \mathcal{V} using the Kleene closure $*$. Then, we may use the following equivalent decision rule

$$(u_1^T, d) \mapsto \hat{u}_{T+1} = \arg \max_{u_{T+1} \in \mathcal{V}^*} \{p(d | u_{T+1}, u_1^T) \cdot p(u_{T+1} | u_1^T)\}. \quad (10)$$

Proof.

$$\begin{aligned} & \arg \max_{u_{T+1} \in \mathcal{V}^*} \{p(u_{T+1} | u_1^T, d)\} \\ &= \arg \max_{u_{T+1} \in \mathcal{V}^*} \{p(u_{T+1} | u_1^T, d) \cdot p(d | u_1^T)\} \\ &= \arg \max_{u_{T+1} \in \mathcal{V}^*} \{p(u_{T+1}, d | u_1^T)\} \\ &= \arg \max_{u_{T+1} \in \mathcal{V}^*} \{p(d | u_{T+1}, u_1^T) \cdot p(u_{T+1} | u_1^T)\} \end{aligned}$$

Here, the first step is obtained by multiplying with $p(d | u_1^T)$, which does not change the argument of the maximization. \square

A.2 Experiment details

For all of our experiments and models, we use BART-large (Lewis et al., 2020a), which adheres to the standard Transformer model from Vaswani et al. (2017) and consists of 12 encoder and 12 decoder layers with a hidden size of 1024 and 406M parameters. We finetune each model for 10 epochs using an initial learning rate of $6.25e - 5$, with no warmup steps, and linear learning rate decay. We use a batch size of 32 by using gradient accumulation. The model is evaluated on the validation set after each epoch and the model with the smallest eval loss is picked as our final model. We truncate the dialog history at 384 tokens and restrict the

length of the grounding to 128 tokens, after which it is cut off.

Except for the Bi-Encoder, which we train using the Triplet Loss, all models are trained using the Cross-Entropy criterion.

All of our experiments using online decoding use $\lambda_1 = 0.6$, $\lambda_2 = 0.4$ and all of the reranking experiments use $\lambda_1 = 0.5$, $\lambda_2 = 0.2$.

All models were trained and evaluated on NVIDIA 1080 or 2080 GPU's.

A.3 Dataset statistics

Dataset	Split	Domain	# Dialogs	#Documents
DSTC9 (Kim et al., 2020)	train	Task-oriented	19,184	2,900
	test		1,981	12,039
Personachat (Zhang et al., 2018)	train	Open-domain	10,907	1,155
	test		1,000	
Wizard-of-Wikipedia (Dinan et al., 2019)	train	Open-domain	18,430	93M sentences
	test_seen		965	93M sentences
	test_unseen		968	93M sentences
FaithDial (Dziri et al., 2022a)	train	Open-domain	4,094	93M sentences
	test		791	93M sentences
Doc2Dial (Feng et al., 2020)	train	Task-oriented	3,474	487
	test		787	572

A.4 Sample model outputs

In this section we present a number of cherry-picked examples for multiple datasets to supplement the qualitative analysis from Section 6.5.

Furthermore, we present examples obtained with different scaling factors to highlight the behavior of the model under these conditions.

Personachat

Grounding <i>u_T</i> Ground truth	i enjoy listening to classical music . i'm a christian . my favorite color is red . i can drive a tractor . my sister is a pole dancer . yes it is . i am not a fan of high maintenance people as well i am far from that , i drive a tractor and do not mind taking it easy
Direct Model	what do you do for a living?
Reranking	do you have any hobbies? i like to drive a tractor
Noisy Channel	i'm not either, my sister is a pole dancer and she's very high maintenance
CTRL	do you have a favorite color? mine is red
CTRL Online	my sister is a pole dancer, and she's very high maintenance
Grounding <i>u_T</i> Ground Truth	i love playing softball . my father is in the army . i recently was fired and had to get a new job . my mom passed away when i was young . i work in the insurance business . my german shepherd is barnaby
Direct Model	did you name him after someone
Direct Model	do you have any pets
Reranking	what do you do for a living
Online Decoding	that's a cute name
CTRL	my mom passed away when i was young
CTRL Online	i work in the insurance business and recently got fired
Grounding <i>u_T</i> Ground Truth	i read twenty books a year . i'm a stunt double as my second job . i only eat kosher . i was raised in a single parent household . wow ! i do love a good horror movie . loving this cooler weather but a good movie is always good .
Direct model	i like the cooler weather too, but i prefer the hot weather.
Reranking	what do you do for a living? i am a stunt double for a movie.
Noisy Channel	i love the cooler weather too, i like to stay inside and read.
CTRL	i have a second job as a stunt double.
CTRL Online	i'm a stunt double as my second job. do you have any hobbies?
Grounding <i>u_T</i> Ground truth	i've a german shepherd named barnaby . i am employed by the us postal service . my father drove a car for nascar . my favorite color is gray . i'm thirty two . i work in the banking industry , boring . i work for the us postal service . it is pretty boring as well .
Direct Model	i work for the us postal service.
Reranking:	that sounds boring. i work for the us postal service.
Online Decoding	i work for the us postal service. do you have pets?
CTRL	i work for the us postal service.
CTRL Online	i am employed by the us postal service, boring too.
Grounding <i>u_T</i> Ground truth	my family has raised horses for a generations . i dream of moving to the city . i'm from a small town . i was raised on a horse farm . i want to be a lawyer . marriage is fun . . we just bought our first house . we built a house on my family horse farm . we have over 5000 acres .
Direct Model	where are you from? i'm from a small town in the midwest.
Reranking	that's great! where are you from? i'm from a small town.
Online Decoding	that's great! i grew up on a horse farm and have always wanted to move to the city.
CTRL	where are you from? i'm from a small town.
CTRL Online	my family has been raising horses for generations. i want to move to the city someday.
DSTC9 Track1	
Grounding <i>u_T</i> Ground Truth	You may bring 2 domestic pets, they must be kept on a short lead, or carried in a basket size 85x60x60cm. So pets of any size can go on the train? Yes, but for pets in a basket, the basket must be no larger than 85x60x60cm. Would you like me to book the train?
Direct Model	Pets must be able to fit into a basket that's 85x60x60cm. Can I help you with anything else?
Reranking	Yes, you may bring your pet on board on a short leash or in a basket 85x60x60cm. Do you have any other questions?
Online Decoding	You can bring 2 domestic pets on a short leash or in a basket 85x60x60cm. Is there anything else I can help you with?
CTRL	Yes, you may bring your pet on board on a short leash or in a 85x60x60cm basket. Anything else I can do for you?
CTRL Online	You may bring 2 domestic pets, but they must be kept on a short lead or carried in a basket 85x60x60cm. (...)
Grounding <i>u_T</i> Ground Truth	English is spoken at Taylor Hotel San Francisco. Thanks, and do they speak any languages other than English? Only English is spoken at the Taylor Hotel San Francisco. Would you like me to find any more details?
Direct Model	Yes, the hotel does speak English. Would you like to make a reservation?
Reranking	No, they only speak English at the hotel. Would you like me to book a room for you?
Online Decoding	No, they only speak English at the hotel. Would you like me to book a room for you?
CTRL	No, they only speak English at the hotel. Is there anything else I can help you with?
CTRL Online	No, they only speak English at the hotel. Is there anything else I can help you with today?
Grounding <i>u_T</i> Ground Truth	Yes it can be, as long as they have access to a payment card for identification. Can my son pick up these tickets for me? Yes as long as they have access to the payment card for identification. Is there anything else I can help you with?
Direct Model	Yes, your son can pick up the tickets for you. Anything else I can do for you?
Reranking	Yes as long as they have access to the payment card for identification. Is there anything else I can help you with?
Online Decoding	Yes as long as they have access to the payment card for identification. Is there anything else I can help you with today?
CTRL	Yes as long as he has access to the payment card for identification. Is there anything else I can help you with?
CTRL Online	Yes as long as they have access to the payment card for identification. Is there anything else I can help you with?

WoW

Grounding <i>u_T</i> : Ground truth	She was the founder and namesake of the Chanel brand. - I learned that the founder of the Chanel brand's first name is Coco. Are you familiar with the brand?
Direct Model	Have you ever heard of Coco Chanel? She was the founder of the Chanel brand.
Reranking:	Have you heard of Coco Chanel? She was the founder and namesake of the Chanel brand
Online Decoding:	I love Coco Chanel. She was the founder and namesake of the Chanel brand.
CTRL	She was the founder and namesake of the Chanel brand.
CTRL Online	Coco Chanel was the founder and namesake of the Chanel brand
Grounding <i>u_T</i> Ground truth	The giant panda ("Ailuropoda melanoleuca", "black and white cat-foot"; , "big bear cat"), also known as panda bear or simply panda, is a bear native to south central China. - Hi! I love giant pandas, they are my favourite animal. they're also called "big bear cat" sometimes? how about you?
Direct Model	Have you ever seen a giant panda? They are native to south central China.
Reranking	My favorite animal is the giant panda. They are native to south central China.
Online Decoding	Giant pandas are so cute! They are native to south central China.
CTRL	The giant panda is a bear native to south central China.
CTRL Online	The giant panda is a bear native to south central China.

Doc2Dial

Grounding <i>u_T</i> : Ground truth	Financial difficulties what does my circumstance have to be to get a general forbearance? Are you having financial difficulties?
Direct Model	Do you have financial difficulties?
Reranking	You would have to have financial difficulties
Online Decoding	Financial difficulties are one of the requirements to get a general forbearance
CTRL	Financial difficulties
CTRL Online	Financial difficulties
Grounding <i>u_T</i> Ground truth	you may request a hearing online. Yes, I disagree with the determination you made at the reconsideration level. You may then request a hearing online.
Direct Model	You can request a hearing online.
Reranking	You can request a hearing online if you disagree with the determination.
Online Decoding	In that case you may request a hearing online.
CTRL	You may request a hearing online.
CTRL Online	You may request a hearing online.

Grounding u_T $\frac{\lambda_1}{\lambda_2} = 0.5$ $\frac{\lambda_1}{\lambda_2} = 2.0$	In computing, React (sometimes styled React.js or ReactJS) is a JavaScript library for building user interfaces. I'm not sure I've heard of it before, what is it? It is a JavaScript library for building user interfaces. React (sometimes styled React.js or ReactJS) is a JavaScript library for building user interfaces.
Grounding u_T $\frac{\lambda_1}{\lambda_2} = 0.5$ $\frac{\lambda_1}{\lambda_2} = 2.0$	David Suchet stars as the eponymous detective, Agatha Christie's fictional Hercule Poirot. No do you know what's it about? It's about a detective named Hercule Poirot, played by David Suchet. David Suchet stars as the eponymous detective, Agatha Christie's fictional Hercule Poirot.
Grounding u_T $\frac{\lambda_1}{\lambda_2} = 0.5$ $\frac{\lambda_1}{\lambda_2} = 1.4$	my friends are all skateboarders . i've a broken arm . i am a skateboarder . my favourite skater is rob dryden . oh no ! what is wrong with it ? are you a pro skateboarder i broke my arm trying to skateboard with my friend rob dryden i broke my arm and my friends are all skateboarders
Grounding u_T $\frac{\lambda_1}{\lambda_2} = 0.5$ $\frac{\lambda_1}{\lambda_2} = 1.4$	i like to work on vintage cars . i own two vintage mustangs . i like to go to country concerts on weekends . my favorite music is country . i have two dogs . that's amazing . i'm currently studying spanish . i love to go to country concerts on the weekends i love to go to country concerts on the weekends with my dogs
Grounding u_T $\frac{\lambda_1}{\lambda_2} = 0.5$ $\frac{\lambda_1}{\lambda_2} = 1.4$	i don't drink soda . i've a boyfriend . i live in a house . i like to eat hamburgers and french fries . i try to work and stay busy , i am in between jobs right now . what do you do for a living? do you like hamburgers and french fries? i love them.
Grounding u_T $\frac{\lambda_1}{\lambda_2} = 0.5$ $\frac{\lambda_1}{\lambda_2} = 1.4$	i volunteer my time with a local bunny rescue . my favorite color is navy blue . i work fulltime in a shipyard . i donate a part of my salary to charity each month . wow that's exciting good for you . i'm learning about computers , programming that's cool. i work at a shipyard. i love computers and programming. cool. i work in a shipyard full time. what do you do in your spare time?
Grounding u_T $\frac{\lambda_1}{\lambda_2} = 0.5$ $\frac{\lambda_1}{\lambda_2} = 1.4$	i'm a descendant of christopher columbus . i love to cook paella . i've a weakness for fish and chips . i visit europe twice a year . i am an art major in college . europe must be very beautiful my sister went there it is. i love it. i've been there twice a year since i was in college. it is. i love it. i go twice a year, and i cook paella and fish and chips
Grounding u_T $\frac{\lambda_1}{\lambda_2} = 0.5$ $\frac{\lambda_1}{\lambda_2} = 1.4$	my family has raised horses for a generations . i dream of moving to the city . i'm from a small town . i was raised on a horse farm . i want to be a lawyer . keep it in your family , that's nice for future generations thank you. i want to move to the city and live on a horse farm. thank you. i want to move to the city one day, but i live in a small town
Grounding u_T $\frac{\lambda_1}{\lambda_2} = 0.2$ $\frac{\lambda_1}{\lambda_2} = 1.4$	Opening hours are from 11am to 5pm on Wednesdays only. What time does the Oakland Museum open? The museum opens at 11am on Wednesdays and at 5pm on Sundays. Do you have any other questions about the museum? The museum opens at 11am on Wednesdays only. Is there anything else I can help you with?
Grounding u_T $\frac{\lambda_1}{\lambda_2} = 0.2$ $\frac{\lambda_1}{\lambda_2} = 1.4$	Cathedral of Saint Mary of the Assumption is open from Monday to Friday from 9:30 am to 4:00 pm. On Saturday it is open from 11:00 am to 5:30 pm and on Sunday it is open from 9:00 am to 3:30 pm. Yeah, is it open on weekends too? The Cathedral of Saint Mary of the Assumption is open from Monday to Friday from 9:30 am to 4:00 pm. Is there anything else I can help you with today? It is open from Monday to Friday from 9:30 am to 4:00 pm. On Saturday it opens at 11:00 am to 5:30 pm and on Sunday it is open at 9:00am to 3:30. Is there anything else I can help you with?