COMP5214 Project Report: Improving factual faithfulness for prompt-based knowledge-grounded dialogue in few-shot settings

Willy Chung, Cheuk Tung Shadow Yiu
(whcchung, ctyiuac)@connect.ust.hk

Abstract

In this project, we experiment on few-shot knowledge-grounded dialogue generation while focusing on evaluating if those generations are factually faithful or not by using a recently proposed evaluation metric defined for this purpose. We show that the zero-shot dialogue generation with T0 is able leverage knowledge in both the context provided, and also in the dialogue history to generate a factual output, while not having been pre-trained on any dialogue generation tasks. We further investigate the robustness of the metric towards using other underlying models for different modules, and show that the overall evaluation score is not affected. Our code is available here[1]

1 Introduction

Knowledge-grounded dialogue agents are conversational systems that leverages external textual information provided by the user, such as a supporting document or a webpage, to generate a discussion on the given topic. All state-of-the-art and best performing agents are based on very large language models (LMs) pretrained on considerable amount of text data (Gopalakrishnan et al. (2019), Roller et al. (2021)), that are then further finetuned for the downstream task. By leveraging the internal representation of natural language that those models have learned during pre-training, their ability to generate near-fluent responses became much better. While those dialogue agents seem more natural and human-like in their response, previous work has shown that the generated text is often not faithful to the provided knowledge: in other words those large LMs often tend to hallucinate (Ji et al. (2022), Dziri et al. (2021b)). This raises serious concerns over safety and controllability of such agents as they may answer to questions counterfactually but in very convincing ways.

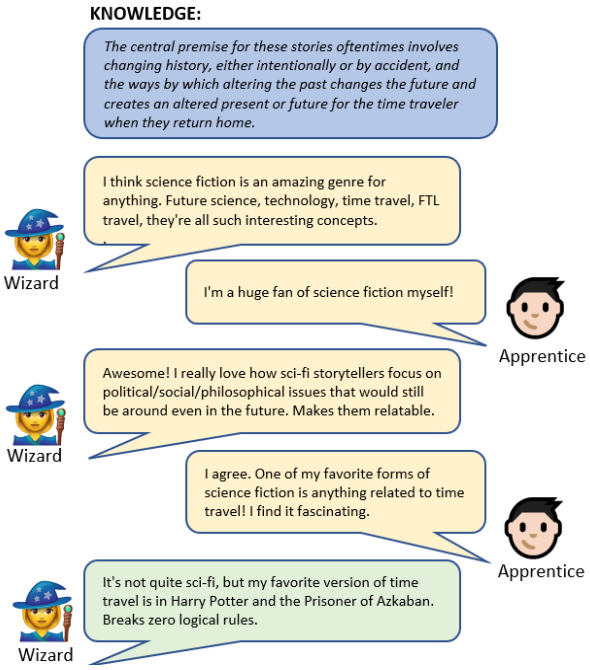

Figure 1: Example of Knowledge-Grounded Dialogue data sample from the Wizard of Wikipedia dataset

Previous research has been done for improving the factual faithfulness of such models mainly through a more controlled training setup (Rashkin et al., 2021), or using external knowledge graphs (Dziri et al., 2021a). However, those methods are very expensive both computationally and timewise. In this work, we explore a way to improve goundedness without any further training for the models, but by leveraging the few-shot capability of the larger LMs (Madotto et al., 2021). Recent work has shown that simply tuning the prompt or providing a few in-context examples can generate as good -if not better- predictions (Lester et al. (2021), Reynolds and McDonell (2021)). More specifically, we propose a framework for few-shot in-context learning in knowledge-grounded dialogue that aims at

[1] https://github.com/WillyHC22/comp5214-groundedness-kgd

Prompt		
Context: "Bananas are very sweet, but they can be the cause of tropical diseases." ***Conversation:*** Amy: "I like bananas." Chris: "Why?" Amy: "Because they are very sweet!" Chris: "I don't like bananas." Amy: "Why?" **<TASK>**		
<TASK>	**Generated output**	**Factual Faithfulness**
Write Chris' answer.	Amy doesn't like bananas because they can be the cause of tropical diseases.	Hallucination
Write a follow up conversation.	Amy likes bananas because they are very sweet. Chris doesn't like bananas.	Neutral
What did Chris say?	Chris doesn't like bananas because they can be the cause of tropical diseases.	Knowledge-Grounded

Table 1: Case illustration of various output using T0_3B given the same prompt, but with different task formulation. Just by changing the task formulation, the output can greatly vary in groundedness

improving the factual faithfulness of the answers by automatically crafting prompts that will guide the model in better understanding the concept of what an answer grounded in knowledge is. Our proposed method consists of crafting the few-shot examples using a mix of extracted unanswerable and answerable questions given a context. The motivation is to provide the model, which already understand natural language to a certain extent, more in-context knowledge about whether a certain information is grounded or not, independently from the fact that the question is well-formulated or not.

However, for seq2seq model, task formulation is very important regarding factual faithfulness. A case illustration is shown in table 1, where different generated output are shown using T0 (Sanh et al., 2021). The prompt is made of a context about bananas and a conversation between two persons on the given context. For different ways to frame the task formulation to the model, the generated output can greatly differ, and more importantly the faithfulness of the output varies from completely counter factual to knowledge-grounded.

In this work, we aim to improve factual faithfulness through prompt engineering in few-shot settings, and alleviate the need of computationally expensive methods which are mostly being focused on recently. We also investigate the robustness of the evaluation metric when the underlying models used are modified, and show that the large multitask transformers are able to leverage knowledge in a zero-shot setting even when they have never seen any knowledge-grounded dialogue generation tasks before.

2 Related work

Hallucination in text-generation has mainly been tackled in tasks such as summarization (Maynez et al., 2020) and news generation, and very recently in knowledge-grounded dialogue generation.

Since one possible reason for hallucination is that the models struggle to put enough importance on the knowledge given, some work tried to alleviate this issue by leveraging knowledge graphs' ability to anchor knowledge in a more structured way. (Dziri et al., 2021a)' neural path hunter follows a generate-than-refine strategy using path grounding through knowledge graphs. The model provides a generated dialogue turn in which possible sources of hallucination are first identified, and then corrected by retrieving the factually appropriate entities through the graph. Their method shows strong result in both reducing hallucinations, but also in identifying them.

A similar approach to improve knowledge selection is to add a retrieval module to help the model select more relevant knowledge. This method follows the retrieval-augmented generation (RAG) architecture, originally proposed to improve open-domain question answering tasks. (Shuster et al., 2021) compares using a dense passage retriever or a re-ranker to help the generation, and shows that the former improves factual faithfulness on top of having the advantage to be more generalizable. This approach, like the previous one, requires to train an additional module on top of the conversational agent.

Lastly, a more recent approach is to directly improve the training of the dialogue model by quantifying the objectiveness and informativeness of the training data first. Through different evaluations, (Rashkin et al., 2021)'s method aims to disentangle the provided information based on how faithful it is, which consequently adds more controllability to the dialogue model's generation through the training process. This information is

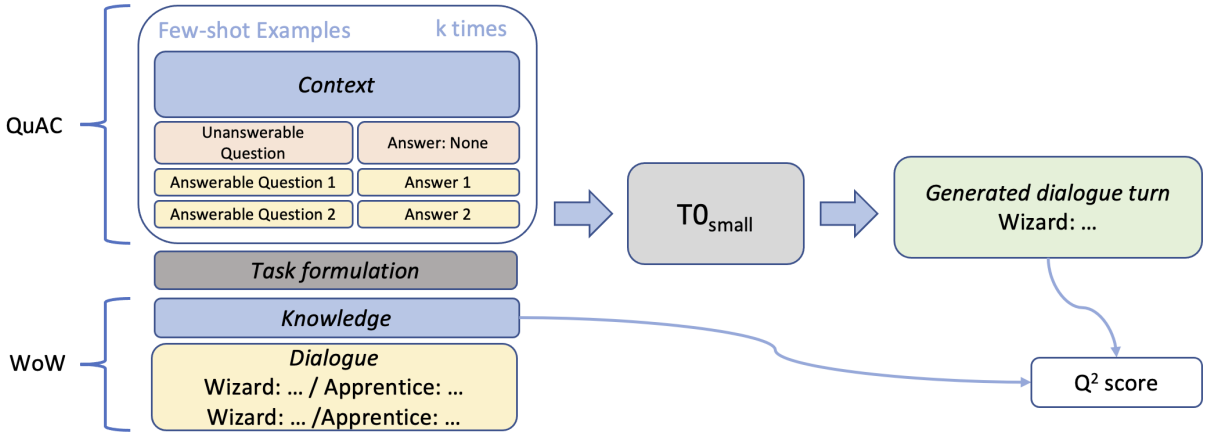

Figure 2: Overall framework and settings for the few-shot

encoded as control features during the training and provided as part of the input. They also investigate applying the same method during decoding, and both shows that it is possible to reduce hallucination through a better training process.

3 Methodology

3.1 Dataset

We need two datasets for this project, one for evaluation and the other one for crafting the few-shot examples. To be coherent with our motivation claim, we have to exclude datasets that have been used during the pre-training of T0 to avoid any bias.

Wizard of Wikipedia (WoW) (Dinan et al., 2018) WoW is a knowledge-grounded dialogue dataset with conversations directly grounded with knowledge retrieved from Wikipedia. This dataset has been originally built to provide both knowledge-grounded and open-domain conversation within the same framework, as most of the knowledge-grounded dialogue datasets usually set explicit roles for the speakers. The dialogue setting is about two participants engaging in chitchat. One will play the role of a knowledgeable expert (the ***Wizard***) while the other is a curious learner (the ***Apprentice***). They use the Transformer Memory Network models, that are capable of retrieving and attending to such knowledge and provide a response, either in the retrieval or generative modes. Because of the input length soft limit of seq2seq models such as T0 (512 input token), we choose to use the shorter conversations of the dataset so that we can provide enough few-shot examples in the input. We will directly test our method on the test set of WoW in the few-shot settings.

QuAC (Choi et al., 2018) For the QA dataset from which we will extract both unanswerable and answerable questions to craft the few-shot, we chose to use the Question Answering in Context (QuAC) dataset for multiple reasons. This dataset provides context for each question that we can interpret as knowledge, and contains a good amount of unanswerable questions. Since we will use both the QA dataset and the knowledge-grounded dialogue dataset in the input prompt, we aimed to get a QA dataset which structure is as close as possible to the dialogue dataset. QuAC is very similar to WoW as the questions are in narrated form between a teacher and a student. We use the training set of QuAC which contains around 10k contexts to craft the few-shot, with more details in section 3.3

3.2 Prompt adaptation

From the QA dataset used to craft the prompt, we define our k-shot examples data $\mathcal{F} = \{(C^k_{QA}, Q^k_{unans}, Q^k_{ans,i}, A^k_i)\}^2_{k=0}$ with C^k_{QA} the context, Q^k_{unans} the extracted unanswerable question, $Q^k_{ans,i}, A^k_i$ the i^{th} extracted answerable question-answer pair for example k. We also define our knowledge-grounded dialogue data $\mathcal{D} = \{C_d, D_d, G_d\}$ where C_d is the context of the dialogue d, D_d the dialogue and G_d the generated dialogue turn. Our prompt adaptation consists of building $\mathcal{F}$ from the QA dataset, and provide the final input sequence for the model as $\{\mathcal{F}, \mathcal{D}\}$. The overall structure of the final prompt given for the few-shot is shown in figure 2

3.3 Dataset processing

To craft the few-shot, we need to decide on several heuristics to process the two datasets discussed in subsection 3.1, with the main limitation being the input size limit of our seq2seq model of 512 tokens. We end up with a prompt structure as shown in the overall framework in figure 2

Few-shot settings We chose to process the training set of QuAC (Choi et al., 2018) to craft the few-shot examples by following the prompt formulation used by the original authors of T0 (Sanh et al., 2021) during pre-training for extractive QA task. To process SQuAD, they reformulated the task as : "Answer the question depending on the context. Context: *context*. Question: *question*. Answer: *answer*". We follow this formulation to integrate the few-shot, with additional heuristic to limit the examples' length as mentioned earlier. We want to extract three questions per context in QuAC, two answerable question and one unanswerable question, where the answer given is CANNOTANSWER. We only consider data sample which contains at least one unanswerable question. For each of those context, we pick the two answerable questions which have answers that appears the earliest in the given context. This allows us to cut the context by more than half on average, as any information given after the second extracted answer is meaningless towards the few-shot. If there is more than one unanswerable questions, we chose to pick the shortest questions, again as a mean to limit the input token limit. Since T0 has not been pre-trained to process a CANNOTANSWER token as unanswerable question's answer, we reframe them as "None" in order to avoid adding more vocabulary to the one existing in the pre-trained tokenizer.

Dialogue generation task As we did for the few-shot setting, we process the training set of the Wizard of Wikipedia dataset (Dinan et al., 2019) to craft the second half of the prompt. Since T0 has not been pre-trained on any knowledge-grounded dialogue generation task, there is no original template for the task formulation to follow. We manually investigated a few task formulations with an example shown in table 1, and decided on the following formulation: "Given the knowledge and the conversation, write the next turn of the conversation. Knowledge: *knowledge*. Conversation: *conversation*" since it was less likely to generate a hallucinated output, while being data agnostic since we do not mention interlocutors directly. Finding a better prompt formulation for dialogue generation will be part of future works since we believe there is significant possible improvement there. As to limit the complexity of the dialogue generation, we shorten the dialogue turns to keep only 4-5 turns, depending on whether the wizard or the apprentice start the conversation. We chose to infer on the wizard's turn, since the dataset originally provides the gold knowledge to leverage for the wizard on the turn he has to answer.

4 Experiment settings

4.1 Language Model

We chose to use T0 (Sanh et al., 2021) as the pre-trained language model to run our experiments. T0 is seq2seq language model based on T5 (Raffel et al., 2020) to assess zero-shot generalization capability of very large language models. During the pre-training of T0, a considerable amount of natural language tasks have been mapped to a prompted form. Both the prompting and the zero-shot generalization makes T0 a good model to use in the context of this work. Furthermore, T0 has not been pre-trained on dialogue generation task but on extractive QA task. While the formulation for both is similar, the dialogue generation task requires a more intricate understanding of dialogue history and turns. Along the analysis for factual faithfulness, we aim to investigate the overall capability of zero-shot dialogue generation for T0. Due to resource limitation, we will use the smaller version of T0 with 3 billion parameters instead of the original one with 11 billion parameters.

4.2 Evaluation

Q^2 Metric (Honovich et al., 2021) The principal metric we use in this work is Q^2 (Honovich et al., 2021), a recently proposed score to evaluate the factual consistency of knowledge-grounded dialogue generation which shows high correlation with human judgement. This metric leverages question generation, question answering and natural language inference to perform the evaluation. Q^2 extracts information spans from the generation which are usually nouns, and generates a question for which the candidate answer is this information span. A follow-up question answering module will answer the generated question using

the knowledge, and both answers will be compared and matched using natural language inference. The generated questions passes a filtering process to identify them as valid or not, if valid then the score of the sample will be 1 in case of entailment and 0 in case of contradiction. If no valid questions are generated, then the fallback will do the inference directly between the knowledge and the generated answer, and add a score of 0.5 in case of neutral generation. For further details, we redirect the reader to the original Q^2 paper (Honovich et al., 2021). We run two experiment settings: one with the question generation done using a pre-trained T5 large for question generation, and the other using the same model we use to generate the dialogue turn, T0 3b, since this part allows more flexibility. For the question answering module, we follow the original author's choice of using Albert-XLarge (Lan et al., 2020). While this is quite expensive to compute, there are not many evaluation metrics specifically dedicated for factual faithfulness so far, and it will be subject of further analysis in the following sections along with an analysis on the metric's robustness towards model quality.

Baseline For the evaluation baseline, as we are using the Q^2 score, we chose to re-implement the baseline for consistent and inconsistent dialogue generation provided by the original authors (Honovich et al., 2021) using dodecaDialogue system. We use the same dataset they chose to run the simple baseline, one with simple counterfactual generation which yields low Q^2 score, and the other with high consistency. We would like to propose implementing further baseline as future direction, with some recently proposed benchmark for evaluating factual faithfulness such as BEGIN (Dziri et al., 2021b).

5 Results and Analysis

5.1 Overall results

The results are shown in table 2 for the evaluation using T0 as the question generation model, and in table 3 for the one with the original T5 pre-trained for question generation.

Both show similar result in terms of Q^2 score, the zero-shot yields 0.209 and 0.183 Q^2 score for the settings with T0 and with T5 respectively, which is encapsulated between our lower (inconsistent) and higher (consistent) baseline, while the one-shot and two-shot performs significantly worse in terms of factual consistency compared to the zero-shot. This point will be analysed and discussed in a further section 5.3. Since we modify the question generation module to analyze how robust this evaluation is towards the metric, we can see that the number of valid questions generated vary greatly between both settings, with an average of 24.8% of valid questions generated using the zero-shot question generation with T0, and 81.3% of valid questions generated using the pre-trained T5 on question generation, while retaining similar Q^2 score.

	Q^2	Valid questions
Inconsistent Baseline	0.166	20.5%
Consistent Baseline	0.258	21.2%
WoW (Zero-shot)	**0.209**	20.8%
QuAC + WoW (One-shot)	0.023	34%
QuAC + WoW (Two-shot)	0.061	27.6%

Table 2: Results for 0/1/2 shot, with the QG part in Q^2 is done using T0, the same model to generate the dialogue

	Q^2	Valid questions
Inconsistent Baseline	0.139	93.4%
Consistent Baseline	0.526	92.1%
WoW (Zero-shot)	**0.183**	63.7%
QuAC + WoW (One-shot)	0.032	85.4%
QuAC + WoW (Two-shot)	0.083	72%

Table 3: Results for 0/1/2 shot, with the QG part in Q^2 is done using a pre-trained T5 for QG.

5.2 Robustness of the evaluation metric

As shown in both tables 2 and 3, changing the underlying question generation module impacts the number of valid questions generated, as expected, but only very slightly the actual Q^2 score. The only significant difference is that the upper baseline has been pushed even higher when using the original T5 for question generation, but both exhibits similar behaviour for the few-shot: the zero shot is shown to be able to leverage some knowledge for the generation, and both one-shot and two-shot performs significantly worse, with the two-shot being better that the one-shot in both cases.

As discussed by the original author, the heuristics defined in this evaluation metric seems to be robust towards model quality: using a pre-trained model especially for question generation does not significantly impact the direct scoring

Context:	"A zombie (Haitian French: "",) is a fictional undead being created through the reanimation of a human corpse."
Conversation:	Wizard: Have you seen the TV series " The Walking Dead" an American post-apocalyptic thriller." Apprentice: Yes, but it's garbage." Wizard: I like the actor Andrew Lincoln. He plays the lead character as a sheriff's deputy" Apprentice: He does a decent job. He has been meme'd to death because of that role."
Generated	Wizard: I've heard that the zombies are a lot more dangerous than they look
Gold	Wizard: I liked it when he awakens from a coma to find the world that he new is now overrun by zombies.

Table 4: Example of a generated dialogue turn in zero-shot settings compared to the gold answer.

in comparison to using a different model for this module. The advantage of switching this model is to reduce computation cost, as in our case, we re-use the same model for both the dialogue generation and the question generation in the evaluation. Another point that could be argued is that T0 has been pre-trained on so many tasks that he learnt to generate question to a reasonable extent, even though he has not been explicitly pre-trained on any question generation task. This point is discussed in the original T0 paper (Sanh et al., 2021).

5.3 Quality of one and two-shot

Both one-shot and two-shot performs way below our inconsistent baseline, showing that the evaluation of factual faithfulness in this case is misplaced in the first place since the generation itself is most often nonsensical. We believe that the reason for this is that our dialogue generation model (T0) allocates too much importance in the few-shot example crafted from QuAC in the input prompt. We can see this by manually checking some of the generation and observe that a majority of them are actually related to the QA part rather than the dialogue generation part. This might not be surprising, as it is difficult to incorporate multitask inside a single prompt, and it will require further research to explore ways of doing this. However, those two experiments still mainly provide two valuable takeaways. The first one is that the percentage of valid questions seems to be negatively correlated with the Q^2 score. As shown in both tables, while both baselines have very close percentage of generated questions, all the few-shot experiments exhibits much higher variance in this aspect. On top of that, the one-shot, which has the highest amount of valid questions, yields the lowest score. This can be seen as a flaw of the Q^2 metric heuristic's definition of a "valid" question. The other observation is that our two-shot performs better than the one-shot. We would have liked to experiment further, but both the input limit of the model and the resource limitation did not allow us to do that. For a possible direction, we may try to see if higher few-shot examples continues to increase the Q^2 score by using a model that allows longer input sequences like Longformer (Beltagy et al., 2020).

5.4 Zero-shot knowledge incorporation

We can see from the result that the zero-shot is actually able to leverage knowledge to some extent: the Q^2 score is obviously lower than the upper baseline as it is still a zero-shot, but better than the inconsistent baseline. When the input prompt is not corrupted by too much off-topic information as we tried to do for one and two-shot, the model is able to leverage both the knowledge and the dialogue history to some extent to do its generation. As shown in the example given in table 4, since our model is not finetuned on any dialogue generation, we can see that the generated answers diverges from the gold. However, the principal objective is to evaluate the factual faithfulness of the generation towards the knowledge and the conversation given. We can see that the generated answer is able to be both factual towards the given knowledge about the definition of a zombie, and follows the conversation in a logical way.

6 Future works

We see this project as an exploration for few-shot dialogue generation with the recent seq2seq large language models, as well as an overview of factual faithfulness evaluation in this context. We have raised several questions and directions for future works in the prompt crafting, the handling of the input length and further zero-shot analysis. Notably, the two main follow-up to that could be interesting are: in the few-shot prompting, explore better methods to attribute different weight importance to the input prompt, or to handle multitask in a single prompt, and further analysis

on zero-shot dialogue generation and the factual faithfulness of those generations given different models and different settings.

In the prompt formulation, we have arbitrarily picked the amount of unanswerable/answerable question to be 1:2, because we believe that the answerable question handles more knowledge than the unanswerable ones. This can be part of further analysis and experiments with different possible ratios. For the task formulation itself, there are a lot of ways to handle it differently. For the question answering part, we followed the original T0 prompt formulation as to not deviate too far from the original understanding of the task, but this can also be subject to more investigation. Finally, we could add human evaluation on our current generations to see if the factual consistency metric chosen here still provides strong correlation.

7 Conclusion

The proposed method in this project shows that zero-shot knowledge grounded dialogue generation is possible for the large language models that have not been pre-trained on any similar tasks. While the one-shot and two-shot are underperforming, adding more examples seems to improve the generation and should be investigated to higher number of examples. We further explore a way to evaluate factual faithfulness in dialogue generation settings, which has not been explored much so far. The recently proposed metric Q^2 does show strong robustness towards model quality for the question generation, and can still be subject to further analysis.

References

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

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.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *CoRR*, abs/1811.01241.

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, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. 2021a. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. *arXiv preprint arXiv:2104.08455*.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2021b. Evaluating groundedness in dialogue systems: The begin benchmark. *arXiv preprint arXiv:2105.00071*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. *Proc. Interspeech 2019*, pages 1891–1895.

Or Honovich, Leshem Choshen, Roee 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.

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 arXiv:2202.03629*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. 2021. Few-shot bot: Prompt-based learning for dialogue systems. *arXiv preprint arXiv:2110.08118*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

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.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 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.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

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.