Towards Building Coherent and Faithful Conversational Models

by

Nouha Dziri

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Computing Science
University of Alberta

© Nouha Dziri, 2023

Abstract

Dialogue systems powered by large pre-trained language models exhibit an innate ability to deliver fluent and natural-sounding responses. Despite their impressive performance, these models fail to conduct interesting and consistent exchanges of turns and can often generate factually incorrect statements, known as "*hallucinations*", impeding their widespread adoption in real-world applications. These issues seem not to be rectified by simply training autoregressive neural models on a massive amount of Web data and then fine-tuning on a specific dialogue benchmark. Progress towards models that do not exhibit these issues requires evaluation metrics that can quantify their prevalence. Unfortunately, there is a significant progress in models' architectures without a significant progress on how they are being evaluated. What is more, current metrics capture mostly surface-level improvements (e.g., human-likeness) and fail dramatically at measuring a deep understanding of attribution.

This dissertation aims at building coherent and faithful conversational models by addressing existing problems from three perspectives: modelling, data and evaluation. First, I introduce DEMI, a new objective function, which aims to make responses more coherent, interesting and diverse. DEMI focuses on maximizing mutual information between past and known future utterances of a particular turn. This is done by applying the chain rule on mutual information and bounding each term separately. Second, I present Neural Path Hunter (NPH), which follows a generate-then-refine approach by augmenting conventional conversational models with an additional refinement stage enabling them to correct potential hallucinations by querying a knowledge graph. Third, I introduce the BEGIN benchmark designed to evaluate attribution in

knowledge-grounded dialogue systems. Through a comprehensive evaluation study on BEGIN, I show that a broad set of existing automatic metrics do not reliably distinguish attributable abstractive responses from unattributable ones, and perform substantially worse when the knowledge source is longer. And lastly, I discuss the origin of hallucinations in conversational models and link that to noise in dialogue benchmarks and to modelling weaknesses. To address this problem, I follow a data centric approach and introduce a new benchmark, FaithDial, which drastically enhances faithfulness and other dialogue qualities.

Overall, in pursuit of building trustworthy conversational models that can be readily adopted in real-world applications, the present thesis highlights (1) how to embed human-like conversational properties in responses (2) how to make responses more faithful and less hallucinated (3) how to reliably evaluate faithfulness.

Preface

This thesis is the outcome of my own work. I ran all the experiments reported in all chapters. In [1], I have equal contributions with Alessandro Sordoni and Hannes Schulz. However, I have excluded their contributions and included only mine in chapter 2. Overall, this thesis features content based on the following papers:

[1] Sordoni Alessandro*, Dziri Nouha*, Schulz Hannes*, Gordon Goeff, Phil Bachman, Remi Tachet. Decomposed mutual information estimation for contrastive representation learning. In International Conference on Machine Learning 2021 (*ICML*) Jul 1 (pp. 9859-9869). PMLR.

[2] Nouha Dziri, Andrea Madotto, Osmar Zaiane, Joey Bose. Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (*EMNLP*) (pp. 2197-2214).

[3] Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. 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 (*NAACL*): Human Language Technologies, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

[4] Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, Siva Reddy; FaithDial: A Faithful Benchmark for Information-Seeking Dialogue. Transactions of the Association for Computational Linguistics (TACL) 2022; 10 1473–1490. doi: https://doi.org/10.1162/tacl_a_00529

[5] Nouha Dziri, Hannah Rashkin, Tal Linzen, David Reitter; Evaluating Attribution in Dialogue Systems: The BEGIN Benchmark. Transactions of the Association

for Computational Linguistics (TACL) 2022; 10 1066–1083. doi: https://doi.org/10.1162/tacl_a_00506

To my family

Acknowledgements

Writing this acknowledgement was not an easy task for me as it was alluded to by many friends. Every time I decide to write, my mind wanders to the last six years of my graduate studies journey. I still remember when I started my MSc in 2016, I did not know what 'NLP' stands for until I googled the term to barely find a few blog posts and videos. Now looking back, my mind is blown away by the sheer amount of resources available to learn about NLP ranging from papers, tutorials, online courses, workshops to conferences, etc. During this journey, I have grown both academically and personally, and I feel indebted to the people who played an immense role in the achievements I made. I wish to thank them through these lines.

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Osmar Zaiane, for his invaluable guidance. Besides being a talented professor, he is an incredible human being. He has never thought twice to dedicate his time for me whenever I was in doubt. I thank him for his patience, invaluable support and for giving me the freedom to pursue the research problems I was passionate about.

Next, I dedicate this thesis to my role model mom, caring dad, affectionate sister and to my extraordinary extended family. I'm eternally indebted to them. Their unconditional love has given me grit and strength to continue pushing the boundaries when I felt in hardship. They never hesitated to support my decisions and always encouraged me to follow my passion. I must also mention how much they sacrificed for my education, the hardest of which was being separated thousands miles away and not being able to see me for several months or years during Covid.

I am also indebted to my wonderful friend, confidante and collaborator Ehsan

Kamalloo. Ehsan has been my collaborator since the start of my research journey. He is undoubtedly an intelligent researcher. His attention to details and creative thinking impress me during our weekly brainstorming. I have been inspired by Ehsan's high standards, academic integrity, and passion for research. He always faces roadblocks with a positive spirit and his cheerful personality was the greatest mood lifter when I was feeling down.

I should also mention that I'm grateful for the intellectually stimulating environment at Amii, Google Research, Microsoft Research and Mila. I have learned immensely from the daily discussions and talks that I attended almost daily. I am incredibly grateful to Dr. Siva Reddi and Dr. Alessandro Sordoni whose mentorship was critical in advancing my research. I cannot express how much I owe them for their guidance that was instrumental to my development as a researcher. I would like also to thank Eva Schlinger, Tom Kwiatkowski, Diyi Yang, Livio Baldini Soares, Hannah Rashkin, Kristina Toutanova, Dipanjan Das, David Reitter, Tal Linzen, Hannes Schulz, Shikhar Sharma, Geoff Gordon, Edoardo Ponti, Mo Yu, and Sivan Milton.

I wish to thank all my other wonderful friends who throughout my graduate studies, supported me, and helped me keep my sanity. Julia is a great listener, supportive, cheerful with an amazing positive energy. Esin is super warm, gentle, generous and an amazing confidante. Joey is extremely wise, kind and knows the best restaurants in Montreal. Roberto is funny, cheerful, and inspiring. I learned a lot from them even from the briefest conversations. I wish to thank many other wonderful friends including Ahmed Chaari, Andrea Madotto, Spandana Gella, Tristan Deleu, Andjela Mladenovic, Christos Tsirigotis, Kory Mathewson, Tristan Sylvain, Daniel Mitropolsky, Abhinav Gupta, Mohamed Musbah, Saeed Sarabchi, Sheila Schoepp, Emna Gharbi, Asma Hanachi, Rahma ben Atitallah, and Ikram Hbiri.

Finally, I wish to thank my PhD committee members for taking the time to attend several checkpoints and providing valuable comments throughout the last four years. Dr. Alona Fyshe has been a wonderful mentor and excellent professor. Her course "ML

and the brain" has been eye-opening for me and I learned a lot about human cognitive capabilities and how they map to AI systems. She has been extremely supportive during the entire process. I must also thank Dr. Angel Chang who provided many valuable comments for the various projects I undertook. I also wish to thank my external examiner, Dr. Jackie Cheung for his excellent feedback and his invaluable questions. He was extremely supportive and respectful. Lastly, I would like to thank Dr Colin Cherry, for thoughtful questions and feedback.

Table of Contents

List of Tables

List of Figures

Abbreviations

BEGIN Benchmark for Evaluation of Grounded INteraction.

DA Data Augmentation.

IR Information Retrieval.

KB Knowledge Base.

KD Knowledge Distillation.

KG Knowledge Graph.

LM Language Model.

MI Mutual Information.

MLE Maximum-Likelihood Estimation.

MLM Masked Language Model.

MR Mean Rank.

MRR Mean Reciprocal Rank.

NCE Noise Contrastive Estimation.

NLG Natural Language Generation.

NLI Natural Language Inference.

NLP Natural Language Processing.

NLU Natural Language Understanding.

NPH Neural Path Hunter.

PLM Pre-trained Language Model.

QA Question Answering.

VRM Verbal Response Modes.

WoW Wizard of Wikipedia.

Chapter 1

Introduction

Conversations play a key role in maintaining human well-being. They constitute the most natural way of interacting verbally with each other. While exchanging ideas, sharing opinions and expressing emotions might seem evident in our day-to-day conversation, transferring these capabilities to machines has been a challenging hurdle for researchers so far.

Over the past decade, virtual assistants have exploded in popularity and have become omnipresent in our lives assisting our daily schedules and routines, and the global pandemic has even accelerated their adoptions [139]. The chatbot market was valued at USD 526 million in 2021, and it is projected to reach USD 3619 million by 2030 [181]. Users are quickly becoming comfortable with the idea of interacting with a chatbot: from a simple weather query to playing music on the phone. These systems are not only used at home but they are used for various applications across several end-user industries, such as healthcare [90, 75], education [186, 84], and banking [160, 131]. Commercial virtual assistants such as Amazon Alexa[1], Google Assistant [2], Apple Siri[3] and Microsoft Cortana[4] have become the center of interest of big tech companies given their rising range of capabilities and their huge market profit. This has made the field of conversational AI growing very rapidly, attracting many researchers in the

[1]https://en.wikipedia.org/wiki/Amazon_Alexa
[2]https://assistant.google.com/intl/en_ca/
[3]https://www.apple.com/ca/ios/siri/
[4]https://www.microsoft.com/en-ca/windows/cortana

machine learning and NLP community. Systems are becoming more conversational, more fluent and more "intelligent" and all of this is thanks to recent advances in language understanding with AI [132, 27, 14, 1, 164].

Neural language models [9] have revolutionized Natural Language Processing (NLP) in recent years. They are based on the self-supervised learning approach [122], in which a network is trained on a large corpus of data to predict the next word in the sentence. By doing this at scale, models learn rich representations to recognize an immense array of patterns and abstractions. With more data and computing power, these models are getting better and better every year at producing shockingly fluent text [164], as well as displaying impressive emergent abilities such as few-shot learning [14]. They have achieved impressive success for both Natural Language Understanding (NLU) and Natural Language Generation (NLG) tasks such as sentiment classification [21], semantic similarity [24], machine summarization [83], and machine translation [5]. Moreover, these models are being adopted for dialogue modelling [1, 164, 36, 43]. The common recipe is to train these powerful neural network models on masses of raw dialogue data from existing web corpora. By doing this, chatbots and in particular open-ended chatbots have become more capable than ever in conversing about any topic approaching a fluency that resembles creations from science fiction. This ability has sparked controversies in big tech research labs and academia about whether the Google chatbot LaMDA [164]— short for Language Model for Dialogue Applications— is sentient. This has also attracted the attention of several media outlets and generated headlines across the globe claiming that AI is mastering language[5].

However, despite this success, neural conversational models struggle to respond consistently and continuously, with no apparent gaps between dialogue turns. Building a system that can respond convincingly, while being engaging and informative with no apparent gaps between dialogue turns is a challenging problem that researchers have been striving to solve [145, 38, 156, 148, 96]. Further, these models are often prone

[5]https://www.nytimes.com/2022/04/15/magazine/ai-language.html

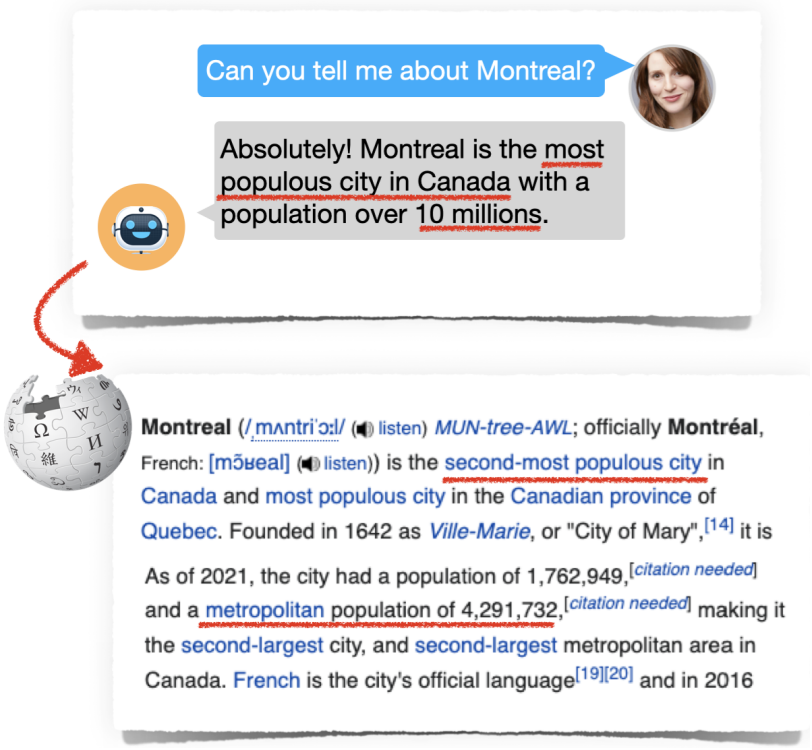


Figure 1.1: An example of a response generated by a neural conversational model. The phrases underlined in red are "hallucinations" not attributable to the Wikipedia article.

to generating unverifiable or factually incorrect statements, a phenomenon known as *hallucination* [135, 39, 150]. They have a disturbing tendency to just make things up out of nowhere without being grounded on truthful facts. Another challenging task of building dialogue systems lies in evaluating the quality of the responses. Models' architectures have significantly changed without a significant progress on how they are being evaluated [53]. Current metrics capture mostly surface-level improvements (e.g., human-likeness) and fail at measuring a deep understanding of faithfulness.

1.1 Thesis Statement

The overarching goal of this thesis is to endow machines with a human-like propensity to converse coherently, engagingly, and truthfully in a highly dynamic environment. Throughout the different chapters, I focus on introducing principled solutions towards this goal. Below, I explain the key aspects of each problem:

1.1.1 Dull, Repetitive and Incoherent Dialogue Responses

Neural response generation approaches rely on powerful auto-regressive language models trained by predicting the next word given the past of the conversation on large corpora of conversational exchanges [161, 132, 170, 68]. Generating utterances from these models can be done efficiently by ancestral sampling, i.e. by sampling words sequentially conditioned on previous words. However, it has been shown that this paradigm typically leads to generic responses, containing too many high frequency tokens and hardly bear any informative words [148, 95]. Although these responses are grammatically correct, they lead to dull and problematic conversations. Further, generated responses tend to follow an illogical reasoning throughout the conversation by producing self-contradictory responses. For example, a neural dialogue system can respond to the utterance "Do you like animals?" by "Yes, I have three cats", thereafter replies to "How many cats do you have" by "I don't have cats.". Therefore, relying only on the Maximum-Likelihood Estimation (MLE) objective function— which maximizes the probability of the next word in the response— is unsuitable as it hardly correlates with the goal of generating consistent and informative responses. Humans are not attempting to maximize text probability when they converse but they are instead trying to achieve goals [56].

> *How can conversational models imitate human-level linguistic capabilities? Can we augment the MLE objective function with a new loss that learns to generate coherent and engaging responses in a self-supervised fashion?*

1.1.2 Hallucination in Dialogue Responses

Despite being fluent, conversational models are still unable to maintain a truthful conversation and may instead hallucinate factually invalid information. The extremely fluent text creates credible impression of human-likeness and users may end up trusting information that are fully hallucinated. Consider the example in Figure 1.1; when the

chatbot is asked information to tell me about Montreal, it says that "Montreal is the most populous city in Canada with a population over 10 millions". By looking into the Wikipedia article, we can notice pretty quickly that the response is wrong. This issue is particularly salient for knowledge-grounded dialogue systems, which are expected to interact with a user in an open-ended fashion while conveying information that is attributable to external identifiable sources. The dangerous part is that these systems can be used by malicious groups to enable large-scale disinformation and inflict harm on members of the society, and marginalized communities always get hit the hardest.

> *It is not yet well-understood why conversational models have a propensity to hallucinate; is it because conversational benchmarks are noisy and contain factually incorrect sentences or does it stem from potential shortcomings in models' architectures and/or training procedures? How can we make existing conversational models faithful? How can we avoid low-quality data releases and build hallucination-free dialogue benchmarks?*

1.1.3 Dialogue Evaluation

In synergy with truthful and coherent dialogue models, evaluating the quality of responses is equally important. Despite advances in dialogue modelling [135, 150, 39], the lack of comprehensive studies on automated evaluation metrics and the lack of testing benchmarks for dialogue evaluation continue to impede progress. Recently, Rashkin *et al.* [136] introduced a human annotation framework that evaluates the attribution of generated text (such as dialogue responses) to a given piece of evidence. An attributed response is one that is connected to a piece of evidence that supports the entirety of the response. Another line of work [70, 39] developed automated metrics to measure attribution[6] of dialogue responses and have reported remarkably high correlations with human evaluations, paving the way for potential alternatives to expensive human evaluations.

[6]Also sometimes referred to as faithfulness in the literature [17, 32, *inter alia*]

However, these approaches suffer from three limitations. First, they evaluate the performance of the metrics on small-scale test data that consist of hundreds of manually annotated examples. This results in high variance and high error on challenging test examples that contain novel patterns not covered in the test set [54]. Second, these metrics essentially ignore the basic elements of a natural-sounding conversation [59, 158, 15]—e.g., backchanneling, acknowledgment, etc—and penalize viable responses that do not convey any specific information, referred to as *generic* responses, such as "*Sorry, I'm not sure about this topic*" or "*What's your favorite food?*". Generic responses are still more desirable than unattributable responses[7] in the context of information-seeking dialogues. In real-world scenarios, it is preferable for a model to acknowledge its ignorance instead of producing hallucinated content. Third, simply looking at correlation with human scores may not sufficiently determine the efficacy and robustness of an evaluation metric as these metrics can be susceptible to spurious correlations, and therefore, may fail at measuring attribution in challenging cases.

> *How can we measure hallucinations effectively? Are state-of-the-art evaluation metrics robust at measuring faithfulness or do they rely mostly on spurious correlations such as word overlap? How can we validate the performance of these metrics?*

1.2 Key Contributions

The main claim of this dissertation revolves around building information-seeking conversational models that are informative, coherent, faithful and engaging. Overall, the desiderata for building such conversational systems can be summarized as follows:

D1: *Coherent, fluent and interesting* conversational models that mirror human conversational capabilities.

[7] Also known as hallucination responses in the literature [39, 150, *inter alia*].

D2: *Hallucination-free* conversational models that can be reliably deployed in real-world applications.

D3: *Reliable evaluation tools*, including testing benchmarks and metrics, to accurately measure the faithfulness of models and to offer ample diagnostic tools when issues arise.

D4: *Faithful conversational benchmarks* that can be used to train models to detect hallucinations and to generate hallucination-free responses.

I tackle the problems from three perspectives: modelling, data and evaluation. My contributions all constitute necessary steps for building systems that are robust in real-world scenarios. In summary, the key contributions of the present work are:

1.2.1 Improving Dialogue Responses by Mutual Information Maximization

Conversation between humans is an ambiguous and complex joint activity. Each utterance is not independent of one another but is instead grounded within a larger dialogue context known to both parties. Modeling the abstract mechanisms that underpin the unique conversational abilities of humans in dialogue systems is a grand challenge especially in the absence of explicit supervision. For instance, it is unclear how to inject current deep learning models with an exhaustive set of rules that mirror human linguistic capabilities. To narrow this chasm, I take inspiration from the self-supervised learning paradigm with the goal of learning conversational characteristics in an implicit fashion (chapter 2). In [155], I propose the idea of maximizing Mutual Information (MI) $I(x; y)$ between the past utterances x and the future utterances y in a given dialogue. I argue that a good dialogue system should be one that estimates representations of the past and the future utterances such that they are mutually predictive. Under this assumption, I decompose $I(x; y)$ by applying the chain rule on MI to obtain a sum of conditional and unconditional MI terms which we call DEMI for DEcomposed Mutual Information. Each term contains smaller chunks of the total

MI that can be approximated with less bias by contrastive approaches. This way, our dialogue model captures long-term dependencies that cannot be explained by the most recent utterance. Quantitative and qualitative results show that DEMI representations result in higher quality responses compared to the baselines. Generated responses are more informative, more diverse, and contained less repetitive content.

1.2.2 Reducing Hallucination

Origin of Hallucination The common belief in the literature is that researchers need to fix the models in order to fix hallucinations. In chapter 5, I investigate both existing benchmarks and generated responses of prominent dialogue models to shed light on the origins of hallucinations [40]. In-depth understanding of the various sources of hallucination and how they manifest themselves can help researchers enforce faithfulness in dialogue models. I take a step closer to gain such an understanding via a systematic study where human evaluators identify and categorize hallucinations in the widely-used benchmarks, measure their frequency, and overall negative impact on generated responses. Analysis revealed that more than 60% of the responses were hallucinated in the datasets, with major hallucination modes that manifest principally through the expression of subjective information (e.g., thoughts, beliefs, feelings, intentions, etc) and the expression of unsupported objective factual information. Similarly, to understand if neural dialogue models make this hallucination more severe, evaluators additionally annotated responses generated by several state-of-the-art models, including ones that are designed to alleviate hallucinations. I find that models do not only hallucinate but even amplify the hallucination behaviour at test time. Overall, I show that hallucination is not only a reflection of training data issues, but also a consequence of the weaknesses of models.

Modelling Neural dialogue models are not necessarily designed to generate faithful outputs, but to mimic the distributional properties of the data. The presence of even

few hallucinated responses may skew the data distribution in a way that curbs the model's ability to generate faithful responses. To address the hallucination challenge from a modelling perspective, I propose Neural Path Hunter (NPH) [39] which focuses on reducing hallucination of neural dialogue models to known facts supplied by a KG (chapter 3). NPH follows a generate-then-refine strategy whereby a generated response is amended using the KG. It leverages a separate token-level fact critic to identify plausible sources of hallucination followed by a refinement stage that retrieves correct entities from a k-hop subgraph. Empirical results show that NPH is capable of reducing hallucination when paired with a number of base dialogue models with relative improvements of 40% over a strong baseline according to human judgements.

Data Even if we come up with a model that is robust enough against hallucination, it will be ultimately bounded by the data quality. To address this problem, I adopt a data-centric solution and create FaithDial [43], a new benchmark for hallucination-free dialogues, by editing hallucinated responses in an existing dialogue benchmark (Wizard of Wikipedia (WoW)) (chapter 5). I observe that FaithDial is more faithful than WoW while also maintaining engaging conversations. I show that FaithDial can serve as training signal for: **i**) a hallucination critic, which discriminates whether an utterance is faithful or not, and boosts the performance by 12.8 F1 score on the Begin benchmark compared to existing datasets for dialogue coherence; **ii**) high-quality dialogue generation. I benchmark a series of state-of-the-art models and propose an auxiliary contrastive objective that achieves the highest level of faithfulness and abstractiveness based on several automated metrics. Further, I find that the benefits of FaithDial generalize to zero-shot transfer on other dialogue datasets. Finally, human evaluation reveals that responses generated by models trained on FaithDial are perceived as more interpretable, cooperative, and engaging.

1.2.3 Dialogue Evaluation

In this dissertation, I focus on evaluating hallucination [43] and attribution [42] in dialogue models. I characterize the consistency of dialogue models as a Natural Language Inference (NLI) problem where I cast a generated response as the hypothesis and the conversation history as the premise. The goal is to understand whether the premise-hypothesis pair is entailing, contradictory, or neutral. While the results illustrated a reasonable correlation with human judgement, the approach was not efficient in detecting hallucination. To counter this, in chapter 5, I introduce a hallucination critic which discriminates whether a response is hallucinated or not. Given the lack of testing benchmarks and the lack of comprehensive studies on automated evaluation metrics, I look into different evaluation challenges and propose the BEGIN benchmark [42] (chapter 4). BEGIN is a challenging, large-scale testing benchmark, for meta-evaluation of grounded dialogue evaluation techniques, comprised of 12k dialogue turns. The main goal of BEGIN is to assess the attribution of model-generated responses with respect to some external knowledge. Based on BEGIN, I investigate the robustness and the reliability of state-of-the art evaluation metrics. I analyze eight evaluation metrics and found that these metrics rely on spurious correlations, do not reliably distinguish attributable abstractive responses from unattributable ones, and perform substantially worse when the knowledge source is longer.

1.3 Dissertation Layout

This dissertation is organized into 6 chapters. Each chapter is a piece of a puzzle that covers one aspect of my ultimate goal — building faithful and coherent conversational models. After the introduction in chapter 1, I discuss the new bound DEMI (chapter 2) which encourages dialogue systems to be more coherent, diverse and informative. chapter 3 concerns fixing entity-based hallucinations in generated responses by retrieving

the correct entities from a KG. Chapter 4 introduces an evaluation benchmark and presents the performance of a large-scale analysis on state-of-the-art evaluation metrics. Chapter 5 studies the underlying roots of hallucinations in conversational models and proposes a new hallucination-free dialogue benchmark. Finally, Chapter 6 summarizes the contributions and discusses potential future research avenues.

Chapter 2

DEMI: Generating Dialogues by Mutual Information Maximization

Recently, natural language understanding and generation models have been successful in solving a variety of natural language processing tasks, such as solving textual entailment, machine translation and question answering. When it comes to modeling open-domain dialogue, however, current systems still lag far behind human capabilities in producing coherent and consistent responses. Generated responses tend to follow an illogical reasoning throughout the conversation by producing either self-contradictory responses or contradicting commonsense facts [176, 37]. In this chapter, we introduce a new bound, DEMI, which aims at improving the quality of dialogue responses using mutual information maximization.

2.1 Introduction

Neural response generation approaches are trained to predict the next word in the sentence given the history using the cross-entropy objective function [161, 132, 170, 68]. However, such a paradigm typically leads to dull, repetitive responses that carry little information [148, 95]. Methods that introduce entropy in the sampling mechanism to induce more diversity have recently been proposed [68, 48]. It remains a problem that sentences containing word repetitions and artefacts that diverge from the statistics of natural language have higher likelihood under the model itself. This suggests a poor

fit to the data distribution. Welleck *et al.* [175] and Li *et al.* [99] specifically train the model to penalize repetitions during training by *unlikelihood* training. However, these methods have a heuristic flavor[1]: they are confronted with the problem of how to choose which words not to repeat and therefore usually degrade performance of the model on metrics like perplexity.

To overcome this issue, we took inspiration from the self-supervised learning paradigm to learn implicitly human conversational characteristics. The objective in self-supervised representation learning approaches is not to maximize likelihood, but to formulate a series of (label-agnostic) tasks that the model needs to solve through its representations [122, 28, 55, 67]. We borrow from the maximum predictive coding framework [44, 3, 108, 123] and argue that a good model for dialogue should be one that learns representations so as to maximize mutual information (MI) between the past utterances and the future utterances $I(x; y)$. The intuition is that when we try to communicate a set of information between each other, we aim to maximize information with respect to the things we deem important and minimize information with respect to the things that we consider less important. The idea is to see what information we can extract from the past and the future of the dialogue that is highly correlated. We note that the maximum-likelihood next-word prediction loss can be considered as a particular form of MI maximization between past and future, where the future is just considered to be the next word[2]. This effectively measures MI between past and future in a particular representation space and learns representations that are predictive of the future one word at a time. Therefore, we investigate whether by extending the way in which we measure and maximize MI between past and future, we can estimate better models of dialogue.

Recent self-supervised learning methods can be seen as training an encoder f such

[1]The tendency of words to repeat once they appeared, i.e. their *burstiness*, is an observed phenomenon in language [23].

[2]This can be intuitively seen by observing that the marginal entropy of the next word is fixed and by considering the language model as minimizing an upper-bound on the conditional entropy of the next word given the past. A set of related observations can be found in [86].

that it maximizes the mutual information (MI) between representations $f(\cdot)$ of a pair of views x and y of the same input datum, $I(f(x); f(y)) \leq I(x; y)$[3]. For sequential

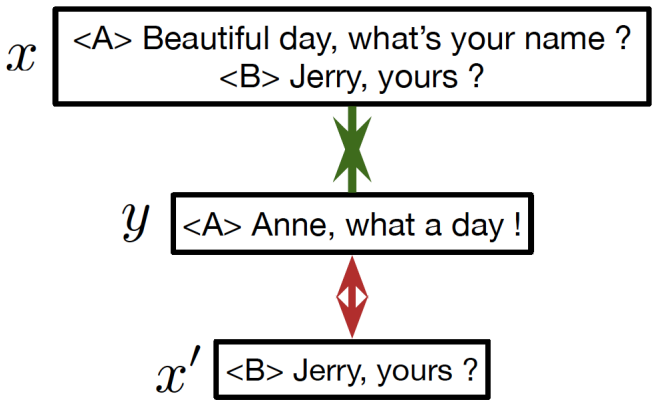


Figure 2.1: A fictional dialogue in which x and y represent past and future of the conversation respectively and x' is the "recent past". In this context, the conditional MI term encourages the encoder to capture long-term dependencies that cannot be explained by the most recent utterances. We can maximize $I(x; y) \geq I(x'; y) + I(x; y|x')$ using a contrastive bound by training x' to be closer to y than to other dialogues from the corpus. Additionally, we train x to be closer to y than to samples from $p(y|x')$, i.e. we can use x' to generate hard negatives y, which corresponds to maximizing conditional MI, and leads the encoder to capture features not explained by x'.

data such as conversational text, the views can be past and future utterances in a given dialogue, or a particular word and its surrounding context [159]. Contrastive approaches train representations of pairs of views to be more similar to each other than to representations sampled from a negative sample distribution. The InfoNCE bound on $I(x; y)$ [123] has been successful insofar as it enjoys much lower variance than competing approaches [153]. However, the capacity of the bound is limited by the number of contrastive samples used [109, 128] and is therefore likely biased when a large amount of MI needs to be estimated.

Given a dialogue, we randomly choose an utterance and consider all the utterances coming before that sentence comprising itself as the past x and those coming after as the future[4] y. We decompose $I(x, y)$ by applying the chain rule on MI to obtain a sum

[3]In what follows, we will slightly abuse language and use the expression "maximizing $I(x, y)$" as a shortcut for "maximizing a lower bound on $I(x, y)$ with respect to f".

[4]We may have considered past/future segments at the token level, but for the purpose of this

of terms, each containing smaller chunks of the total MI that can be approximated with less bias by contrastive approaches. For example, consider creating a subview x' by removing information from x, e.g. by restricting the dialogue history to one sentence as depicted in Fig. 2.1. By construction, $I(x', x; y) = I(x'; y) + I(x; y|x') = I(x; y)$. Decomposed Estimation of Mutual Information (DEMI) prescribes learning representations that maximize each term in the sum, by contrastive learning. The conditional MI term measures the information about y that the model has gained by looking at x given the information already contained in x'. An intuitive explanation of why this term may lead to capturing more of the total MI between views can be found in Fig. 2.1. By setting x' to be the most recent utterance, the encoder is directly encouraged to capture long-term dependencies that cannot be explained by the most recent utterance. Most importantly, the conditional MI term encourages the encoder to capture more non-redundant information across views.

Our model is trained to estimate representations of the past and the future such that they are mutually predictive, i.e. given the past, the true future can be easily distinguished from a set of negative, candidate futures sampled from a proposal distribution. Our contributions are the following: we show that maximizing MI with InfoNCE has a synergistic effect to maximum likelihood estimation for the metrics considered: as a result of the optimization, held-out perplexity decreases. In addition, we demonstrate that DEMI can potentially capture more of the total information shared between the original views x and y. We extend existing contrastive MI bounds to conditional MI estimation and present novel computationally tractable approximations. Supplementally, our results offer another perspective on *hard* contrastive examples, i.e., Faghri *et al.* [46], given that conditional MI maximization can be achieved by sampling contrastive examples from a partially informed conditional distribution instead of the marginal distribution. Our extensive experiments on the Wizard of Wikipedia [30] show that our model is capable of making responses more diverse, coherent and informative

work, we stick to a utterance-level segmentation, which naturally arises in conversational exchanges.

based on automatic metrics and human judgement.

2.2 Problem Setting

The maximum MI predictive coding framework [108, 123, 67] prescribes learning representations of input data such that they maximize MI between inputs and representations. Recent interpretations of this principle create two independently-augmented copies x and y of the same input by applying a set of stochastic transformations twice, and then learn representations of x and y by maximizing the MI of the respective features produced by an encoder $f : \mathcal{X} \to \mathbb{R}^d$ [4, 19]:

$$\arg\max_f \; I(f(x); f(y)) \leq I(x; y) \tag{2.1}$$

where the upper bound is due to the data processing inequality. Our starting point to maximize Eq. 2.1 is the recently proposed InfoNCE lower bound on MI [123] which trains $f(x)$ to be closer to $f(y)$ than to the representations of other texts drawn from the marginal distribution of the corpus. This can be viewed as a *contrastive* estimation of the MI [123] and has been shown to enjoy lower variance than competing approaches [153].

2.2.1 InfoNCE Bound

InfoNCE [123] is a lower-bound on $I(x; y)$ obtained by comparing pairs sampled from the joint distribution $x, y_1 \sim p(x, y)$ to pairs x, y_i built using a set of negative examples, also called *contrastive*, independently sampled from the marginal:

$$I_{NCE}(x, y|\phi, K) = \mathbb{E}\left[\log \frac{e^{\psi(x,y_1)}}{\frac{1}{K}\sum_{k=1}^{K} e^{\psi(x,y_k)}}\right], \tag{2.2}$$

Usually, ψ is the dot product of the representations after applying an additional transformation g, e.g. an MLP, $\psi(x, y) \triangleq g(f(x))^T g(f(y))$ [19].

InfoNCE has recently been extensively used in self-supervised representation learning given that it enjoys lower variance. However, the bound is loose if the true

mutual information $I(x;y)$ is larger than $\log K$, which is likely when dealing with high-dimensional inputs such as text. To overcome this difficulty, recent methods either train with large batch sizes [19] or exploit an external memory of negative samples in order to reduce memory requirements [20, 167]. These methods rely on uniform sampling from the training set in order to form the contrastive sets.

2.3 Decomposing Mutual Information

When $\mathcal{X}$ is high-dimensional, the amount of mutual information between x and y will potentially be larger than the amount of MI that I_{NCE} can measure given computational constraints associated with large K and the poor log scaling properties of the bound. We argue that we can ease this estimation problem by creating subviews of x and applying the chain rule on MI to decompose the total MI into a sum of potentially smaller MI terms.

By the data processing inequality, we have: $I(x;y) \geq I(\{x^1,\ldots,x^N\};y)$, where $\{x^1,\ldots,x^N\}$ are different subviews of x – i.e., views derived from x without adding *any* exogenous information. For example, $\{x^1,\ldots,x^N\}$ can represent single utterances in a dialog x, or sentences in a document x. Equality is obtained when the set of subviews retains all information about x or if x is in the set.

For ease of exposition and without loss of generality, we consider the case where we have two subviews, x itself and x'. Then, $I(x;y) = I(x,x';y)$ and we can write $I(x,x';y)$ by applying the chain rule for MI:

$$I(x,x';y) = I(x';y) + I(x;y|x'). \tag{2.3}$$

The conditional MI term can be written as:

$$I(x;y|x') = \mathbb{E}_{p(x,x',y)} \log \frac{p(y|x,x')}{p(y|x')}. \tag{2.4}$$

This conditional MI is different from the unconditional MI, $I(x;y)$, as it measures the amount of information shared between x and y that cannot be explained by x'.

Lower bounding each term in Eq. 2.3 with a contrastive bound can potentially lead to a less biased estimator of the total MI. This motivates us to introduce DEMI, a sum of unconditional and conditional lower bounds:

$$I_{DEMI} = I_{NCE}(x'; y) + I_{CNCE}(x; y|x') \leq I(x; y), \tag{2.5}$$

where I_{CNCE} is a placeholder for a lower bound on the conditional MI. Both conditional and unconditional bounds on the MI can capture at most $\log K$ nats of MI. Therefore, DEMI in Eq. 2.5 potentially allows to capture up to $N \log K$ nats of MI in total, where N is the number of subviews used to describe x. This is strictly larger than $\log K$ in the standard I_{NCE}.

2.4 Experiments

Setup In addition to the LM loss, we maximize MI between representations of the past and future utterances in each dialogue, i.e. the predictive coding framework [44, 109]. We consider past and future in a dialogue as views of the same conversation. Given L utterances $(x_1, \ldots, x_L)$, we set $y = (x_{k+1}, \ldots, x_L)$, $x = (x_1, \ldots, x_k)$ and $x' = x_k$, where $(.)$ denotes concatenation and k is randomly chosen between $2 < k < L$. The goal is therefore to imbue representations with information about the future that cannot be solely explained by the most recent utterance x'. For each utterance in a dialogue, we encode the past (i.e., previous utterances) and the future (i.e., next utterances) using a neural network encoder. Then, we train the model such that past and future are close in the embedding space, specifically, closer than other possible future candidates drawn from the marginal future distribution.

2.4.1 Training Objective

Our loss function $\mathcal{L}$ extends the classical next-word prediction loss by maximizing additional mutual information terms between past and future. Although it is explicitly possible to maximize all the terms in $I_{\text{DEMI}}(x, y)$ decomposition, in our experiments,

we restrict ourselves to mainly maximize three mutual information terms: (i) the log-likelihood of the next-word given the past, i.e. $\log p(w|x) \approx I(x;w)$, (ii) the basic InfoNCE bound $I_{\text{NCE}}(x;y)$ and (iii) one of the conditional mutual information terms from our $I_{CNCE}(x;y|x')$. Incorporating more than two terms in the decomposition is straightforward and could be investigated in the future. Our loss is a weighted combination of these terms:

$$\mathcal{L} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[- \lambda_1 \sum_{w\in y} \log p(w|x) - \lambda_2 \, I_{\text{DEMI}}(x,y)\Big] \tag{2.6}$$

where λ_i are hyperparameters and $\sum_{w\in y} \log p(w|x) \approx I(x;w)$ is the the log-likelihood of the next-word given the past.

2.4.2 Models

We evaluate our introduced model against different baselines: **GPT2** is a basic small pre-trained model fine-tuned on the dialogue corpus. **TransferTransfo** [183] augments the standard next-word prediction loss in GPT2 with the next-sentence prediction loss similar to Devlin *et al.* [28]. **GPT2-MMI** follows **MMI-bidi** [95]; we generate 50 responses from GPT2 and then rank them based on a trained backward model $p_{GPT2}(y|x)$. For the **InfoNCE** baseline, we only maximize the unconditional MI between x and y and sample negative futures from the marginal distribution $p(y)$. **DEMI** maximizes conditional MI by recurring to using GPT2 by computing representations of past and future are the state. GPT2 is a generative model therefore we can simply sample a set of negative futures from $p_{GPT2}(y|x')$, that is, by restricting the amount of contextual information GPT2 is allowed to consider. To speed up training, the negative sampling of future candidates is done offline.

2.4.3 DEMI Details

The optimization of the DEMI requires the specification of a critic. Following previous work [123, 67], we implement the critic by a dot product between representations of

the past $f(x)$ and those of the future $f(y)$. We obtain f_x, f_y by running a forward pass with the GPT2 model on the words from the past and the future separately and by taking the state of the last layer of the GPT2 corresponding to the last token in the past and the future respectively.

For all DEMI terms, given the past, the model is trained to pick the ground-truth future among a set of N future candidates. This candidate set includes the ground-truth future and $N-1$ negative futures drawn from different proposal distributions. To compute $I_{NCE}(x;y)$, we consider the ground truth future of each sample in the batch as a negative candidate for the other samples in the same batch. Using this approach, the number of candidates N is equated to the batch size. This ensures that negative samples are sampled from the marginal distribution $p(y)$. To compute the conditional MI bound $I_{CNCE}(x;y|x')$, we sample negative futures $p(y|x')$ by conditioning the GPT2 model on the most recent utterance in the past x'. A sample dialogue from WoW and hard negative examples are presented in Table 2.1.

2.4.4 Experimental Setup

Given memory constraints, all the proposed models are trained with a batch size of 5 per GPU, considering up to three utterances for the future and five utterances in the past. All the models are trained on 2 NVIDIA V100s. The models early-stop in the 4th epoch. We use the Adam optimizer with a learning rate of 6.25e-5, which we linearly decay to zero during training. Dropout is set to 10% on all layers. InfoNCE/DEMI terms are weighted with a factor 0.1 in the loss function. We varied the factor from 0.1 to 1 and 0.1 was chosen based on the best results on the validation set. During inference, we use nucleus sampling [68] with $p = 0.9$ for all models.

2.4.5 Automated metrics

Repetition The word repetition metrics aim at testing the model's performance in generating responses while avoiding artificial repetitions. We employ the repetition

x	A:	I like parachuting or skydiving.
	$\boldsymbol{B}$:	I've never done either but they sound terrifying, not a fan of heights.
	A:	But it is interesting game. This first parachute jump in history was made by Andre Jacques.
	$\boldsymbol{B}$:	Oh really ? Sounds like a french name, what year did he do it ?
	A:	It done in October 22 1797. They tested his contraption by leaping from a hydrogen balloon.
	$\boldsymbol{B}$:	Was he successful or did he kick the bucket off that stunt?
	A:	I think its a success. The military developed parachuting tech.
$y \sim p(y\|x')$	B_{gt}	Yeah nowadays they are a lot more stable and well made.
$y_{1:N} \sim p(y\|x')$	B_1:	That is great. I've been skydiving for days now. How is it ?
	B_2:	Oh I have never flown but I'm glad to know.
	B_3:	I've been dying for it since I was a kid.
	B_4:	Yes, that is why NASA had an advanced mechanics tech for months.
	B_5:	I went parachuting last Sunday and enjoyed it.
$y'_{1:N} \sim p(y)$	B'_1:	I think science fiction is an amazing genre for anything
	B'_2:	Can you imagine the world without internet access ?
	B'_3:	I am just finishing my university course and I will be a qualified pharmacist.
	B'_4:	I don't know how to be romantic. I have trouble expressing emotional attraction.
	B'_5:	I think Krav Maga is a martial art sport. That 's the reason I picked it .

Table 2.1: A sample dialogue between speaker A and speaker B from the Wizard of Wikipedia dataset. The four rows from top to bottom are: (1) x: the "past" dialogue up to utterance k (2) y: the ground-truth utterance for the next turn $k+1$ (3) $y_{1:N}$: future candidates sampled from the "restricted context" future distribution $p(y|x')$. These candidates correspond to the set of **hard** negatives that are closely related to the conversation. (4) $y'_{1:N}$: future candidates sampled randomly from the dataset. We can see that candidates $y_{1:N}$ are semantically close but incoherent w.r.t to the dialogue history as they were conditioned solely on the immediate past utterance x'. However, we can notice that candidates $y'_{1:N}$ are semantically distant from x as they were sampled randomly from the data distribution. The highlighted text in green correspond to the topic of the conversation. Speaker B mentions that they have never done either parachuting or skydiving. B_1 corresponds to the utterance generated based on the restricted context x'. The utterance is on-topic but completely contradictory to what speaker B has said in the past. On the other hand B'_1 is randomly sampled from other dialogues. We can observe that the utterance is clearly irrelevant to the conversation.

metrics presented in Welleck *et al.* [175]: **seq-rep-n**, **rep**, **wrep** and **uniq**. These metrics are defined based on the amount of repetitions in the generations. **seq-rep-n** measures the portion of duplicate n-grams in a generated sequence:

$$\textbf{seq-rep-}n = 1 - \frac{|\text{unique n-grams}(w_{1:N})|}{|\text{n-grams}|} \tag{2.7}$$

where $w_{1:N}$ is the generated utterance. We report **seq-rep-avg** which averages over $n \in \{2, 3, 4, 5, 6\}$. **rep** measures the fraction of tokens that occur in previous tokens, **uniq** counts the number of unique tokens on the validation set.

Distinct-n The metric is derived from Li *et al.* [95]. It is defined as the number of unique n-grams, normalized by the total number of n-grams of tested sentences.

Entropy-n We employ the entropy metric from Zhang *et al.* [196] which aims to fix the problem of frequency difference of n-grams in Distinct-n by reflecting how evenly the empirical n-gram distribution is for each given sentence.

2.4.6 Results

Our experiments are performed on the Wizard of Wikipedia (WoW) [30]. WoW dialogue [30] takes place between a Wizard and an Apprentice. The Wizard is tasked with providing information about a particular topic and the Apprentice, in turn, is expected to seek more information. At each turn of the conversation, the Wizard is presented with passages from Wikipedia and chooses a span from the document—typically one or two sentences—that serves as evidence supporting their response. In total, there are 18430, 967 and 968 dialogues (training/dev/test).

Table 2.2 and Table 2.3 show results on the validation set and test set obtained by 3 pretraining seeds. Generated responses by different models are presented in Table 2.4. The automated metrics indicate that DEMI representations result in higher quality responses. The proposed InfoNCE and DEMI bounds achieve lower perplexity, reduce

Model	ppl	BLEU	H-rel	H-hum	H-inf
GPT2	19.21	7.81	✓	✓	✓
TransferTransfo	19.32	7.5	✓	✓	✓
GPT2-MMI	19.30	6.5	✓	✓	✓
InfoNCE	18.85	8.0	=	✓	✓
DEMI	**18.70**	**8.2**	=	=	=
Human	–	–	✗	✗	✗

Table 2.2: Perplexity, BLEU and side-by-side human evaluation on WoW [30]. H-columns indicate whether DEMI was preferred (✓) or not (✗), or neither (=) at $\alpha = 0.01$.

Model	ppl	seq-rep	rep	wrep	uniq	dist-1	dist-2	BLEU	Ent-4
GPT2	19.24	0.064	0.130	0.132	7393	0.064	0.392	7.75	0.095
TranTransfo	19.33	0.078	0.134	0.132	7735	0.058	0.386	7.52	0.084
GPT2-MMI	19.35	0.070	0.129	0.135	7623	0.052	0.384	7.40	0.092
InfoNCE	18.88	0.065	0.126	0.131	8432	0.065	0.390	7.99	0.107
DEMI	**18.66**	**0.050**	**0.120**	**0.128**	**8666**	**0.070**	**0.405**	**8.10**	**0.108**
Gold	-	0.052	0.095	-	9236	0.069	0.416	-	0.110

Table 2.3: Results for perplexity, sequence-level metric, token-level metrics, BLEU and diversity metrics on the test data of WoW.

next-token repetition and increase the number of unique next-tokens compared to the baselines GPT2, GPT2-MMI and TransferTransfo.

Human Evaluation We also perform human evaluation on randomly sampled 1000 WoW dialogue contexts. We present the annotators with a pair of candidate responses consisting of InfoNCE, DEMI and baseline responses. They were asked to compare the pairs regarding informativeness, relevance to the history and humanness (e.g., fluency), using a 3-point Likert scale [197]. In Table 2.2, we see that overall responses generated by DEMI were strongly preferred to other models but not to the gold response. Bootstrap confidence intervals and p-values (t-test, following [197]) indicate

significant improvements at α=0.01. Detailed results are depicted in Table 2.5, 2.6, 2.7. The "pivot" is always the system given by DEMI. Pairings where the pairwise confidence interval is marked with "*" have a significant difference.

Context	**Generated Responses**
A: I recently took up archery and I think I love it . Do you know anything about Archery ? **B:** Archery is a sport / skill of using a bow to propel arrows. **A:** And a great sport it is . Do you know where archery originated from ?	**DEMI**: Archery originated from ancient Mesopotamia, and is actually practiced internationally in over 200 countries. **InfoNCE**: I am not sure but many traditions and principles arose during the Mesolithic and Paleolithic era. **TransferTransfo** : Yep , you just use it for skill and using it to shoot arrows. **GPT2**: I don't know, but I know that the old French called it archer's art.
A: What is your motivation? **B:** That's a tough question. I 'd say my own progress. I like seeing how much I improve with something. You ? **A:** I am retired now. Are you familiar with the work of Mehr and Meyer, well known psychologists ? **B:** They said that " Motivation is a word that is part of the popular culture as few other psychological concepts are . " Do you agree with what they said ?	**DEMI**: Yes, definitely! You should check out Mark Twain's famous motivational book ! There are theories that point out how important it is to meditate. **InfoNCE**: They said that in psychology research, theories and evidence, participants who enjoyed continuous experience could be more likely to be successful. **TransferTransfo**: Absolutely, I disagree. Are you an author or a commentator? **GPT2**: That's awesome. So , what types of differences exist between "good " motivatin and bad?

Table 2.4: Selected responses from different methods fine-tuned on the Wizard of Wikipedia dataset. DEMI responses are more informative and interesting compared to the baselines.

2.5 Related Works

Self-supervised learning Representation learning based on MI maximization has been applied in various domains such as images [60, 18], words [111, 159], graphs [171], RL [107] and videos [72], exploiting noise-contrastive estimation (NCE) [64], InfoNCE [123] and variational objectives (MINE) [67]. InfoNCE have gained recent interest w.r.t. variational approaches due to its lower variance [153] and superior performance in downstream tasks. InfoNCE however can underestimate large amounts of true MI given that it is capped at $\log K$. Poole *et al.* [128] propose to trade-off

Baseline	DEMI_wins	DEMI_CI	baseline_wins	baseline_CI	pairwise_CI	p
GPT2	0.48726	(0.44, 0.53]	0.28662	(0.25, 0.32]	(0.13, 0.27] *	<0.001
GPT2-MMI	0.65833	(0.6, 0.71]	0.16250	(0.12, 0.21]	(0.4, 0.58] *	<0.001
TransferTransfo	0.46888	(0.43, 0.51]	0.30043	(0.26, 0.34]	(0.09, 0.24] *	<0.001
InfoNCE	0.41711	(0.38, 0.46]	0.36748	(0.33, 0.41]	(-0.03, 0.13]	0.0905
gold_response	0.22679	(0.19, 0.26]	0.54325	(0.5, 0.59]	(-0.39, -0.25] *	<0.001

Table 2.5: Which response is more *relevant* to the history?

Baseline	DEMI_wins	DEMI_CI	baseline_wins	baseline_CI	pairwise_CI	p
GPT2	0.45084	(0.41, 0.49]	0.32636	(0.29, 0.37]	(0.05, 0.2] *	<0.001
GPT2-MMI	0.61734	(0.56, 0.67]	0.18393	(0.14, 0.23]	(0.34, 0.53] *	<0.001
TransferTransfo	0.43617	(0.4, 0.48]	0.35000	(0.31, 0.39]	(0.01, 0.16] *	0.0028
InfoNCE	0.44630	(0.41, 0.49]	0.34515	(0.31, 0.38]	(0.03, 0.17] *	<0.001
gold_response	0.22164	(0.19, 0.26]	0.56608	(0.52, 0.61]	(-0.41, -0.28] *	<0.001

Table 2.6: Which response is more *humanlike*?

Baseline	DEMI_wins	DEMI_CI	baseline_wins	baseline_CI	pairwise_CI	p
GPT2	0.56157	(0.52, 0.6]	0.21444	(0.18, 0.25]	(0.28, 0.42] *	<0.001
GPT2-MMI	0.68750	(0.63, 0.74]	0.12292	(0.09, 0.16]	(0.48, 0.65] *	<0.001
TransferTransfo	0.51931	(0.48, 0.56]	0.24571	(0.21, 0.28]	(0.21, 0.34] *	<0.001
InfoNCE	0.41288	(0.37, 0.45]	0.33580	(0.3, 0.38]	(0.0, 0.15] *	0.0059
gold_response	0.32384	(0.28, 0.36]	0.46624	(0.43, 0.51]	(-0.22, -0.07] *	<0.001

Table 2.7: Which response is more *informative*?

between variance and bias by interpolating variational and contrastive bounds. Song and Ermon [154] propose a modification to InfoNCE for reducing bias where the critic needs to jointly identify multiple positive samples at the same time. Our proposal to scaffold the total MI estimation into a sequence of smaller estimation problems shares similarities with the recent telescopic estimation of density ratio [140] which is based on variational approximations. Instead, we build upon InfoNCE, propose new results on contrastive conditional MI estimation and apply it to self-supervised representation learning.

Dialogue A conversational agent needs to check many boxes to be successful. It should balance between simplicity and detail, stay on topic or change it appropriately, ask questions and answer them and generate fluent text [145]. The starting point is generally training large autoregressive models, such as GPT2 models [132, 197], on a massive amount of Web data and then finetuning them on a specific dialogue dataset [31]. Recent works [175, 145, 78] have proposed to control generation by decoding strategies to increase or decrease the probability of certain words that align with specific dialogues' attributes. These methods are mostly applied at test time, requiring no change to the training method, although See *et al.* [145] also suggest to condition the training on some control features.

A parallel problem seems to be the mismatch between the model's learned distribution and the true data distribution [13, 175, 148]. Sentences that contains repetition artefacts have high likelihood under the model and longer generated text tend to become more and more incoherent. Braverman *et al.* [13] introduce a calibration procedure to fix the entropy amplification problem in text generation models. Starting from the observation that maximum likelihood-trained models generate token-level and sequence-level repetitions much more commonly than in the human training distribution, Welleck *et al.* [175] and Nakamura *et al.* [116] introduce training objectives that explicitly reduce the likelihood of generating frequent and repeated responses on

the token level. Welleck *et al.* [175] demonstrate that their *unlikelihood* loss improves over decoding strategies, such as *nucleus* [68] and *top-k* [48] sampling used in conjunction with maximum likelihood-trained models. Li *et al.* [99] extend the unlikelihood training from language modeling to dialogue generation, where it helps mitigate the high incidence of frequent words and repetitions. These methods are usually heuristic ways to limit repetitions, they are confronted with the problem of how to choose which words not to repeat and usually degrade performance of the model on metrics like perplexity. In contrast to these approaches, our proposed objective DEMI extends standard next-word prediction approaches. It positively encourages the model to retain long-term context specific to the current dialogue, which it uses to distinguish the true future from sampled futures.

Other approaches have utilized mutual information within their training process while some have used it only at decoding time to re-rank better. Li *et al.* [97] and Li *et al.* [95] use mutual information to re-rank generated answers and observed that MI led to diverse responses and improved quality on automated metrics and human evaluation. This objective is not applied during training and therefore does not help in estimating a better model overall. Li *et al.* [98] and Zhang *et al.* [192] turn to reinforcement learning to maximize mutual information between context and generated utterances. Our proposed approach is complementary as it works instead on the representation space and can be trained with self-supervised learning objectives.

2.6 Conclusion

In this chapter, we presented an application of the principle of predictive coding through mutual information maximization to dialogue. We argued that next-word prediction is a particular case of MI maximization between past and future, where the future is a single next-word. Therefore, we extended the MI maximization to take into account richer representations of the future. This has been done leveraging a previously proposed InfoNCE bound and a newly proposed DEMI bound. DEMI is

obtained by applying the chain rule on mutual information and bounding each MI term separately. This allows us to show that a particular type of negative samples, i.e. futures sampled from a restricted past, correspond to maximization of conditional mutual information terms. Our experiments suggest that generalizing the standard next-word prediction loss can be beneficial for obtaining a better fit to the data distribution in a dialogue setting. Coincidentally, this also helps in obtaining responses that contain less word repetitions, a recurring problem in dialogue generation models.

Chapter 3

Reducing Hallucination in Dialogue Systems via Path Grounding

Each utterance in a dialogue is not independent of one another but is instead grounded within a larger dialogue context known to both parties [76, 156, 148, 38]. Indeed, if a response to an utterance fails to be faithful to some given knowledge—i.e. by producing false information—it is uninformative and runs the risk of jeopardizing the entire enterprise of conversation. More precisely, this means that in addition to being fluent, coherent, and diverse, utterances within a dialogue must also be *factually correct*.

The faithfulness of responses is of principal importance when designing dialogue systems that are grounded using auxiliary knowledge such as KG. Despite maintaining plausible general linguistic capabilities, dialogue models are still unable to fully discern facts and may instead hallucinate factually invalid information. In this chapter, we focus on addressing the open problem of hallucination of factually invalid statements in knowledge-grounded dialogue systems where the source of knowledge is a KG.

3.1 Introduction

Empirical evidence for hallucination in Language Model (LM) runs contrary to known studies that these large models are capable of recalling factual knowledge, e.g. entities and relations in a KG [141, 127]. This suggests that this inherent lack of

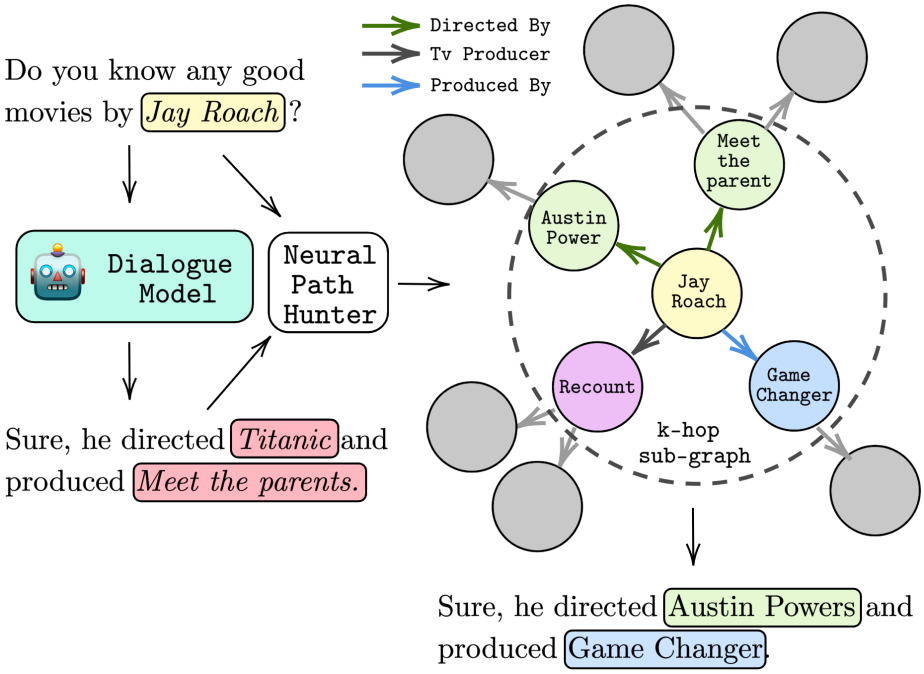


Figure 3.1: Neural Path Hunter overview.

controllability may be remedied by leveraging external oracle knowledge. However, existing approaches to knowledge grounding often suffer from a source-reference divergence problem whereby the reference contains additional factual information and simply training on the reference is insufficient to guarantee faithfulness [182, 126, 165]. Consequently, ensuring the faithfulness of knowledge grounded dialogue systems—via precise alignment of the source and reference—remains an open challenge.

In this chapter, we first identify prominent modes of hallucination by conducting a systematic human study on generated responses which reveals one major source of hallucination as the (mis)-use of wrong entities to describe factual content [87], a problem that persists when naively applying language models in dialogue systems. To enforce faithfulness to the misattribution of entities in grounded dialogue systems, we introduce Neural Path Hunter (NPH), a module that operates on hallucinated responses. NPH follows a generate-then-refine approach by augmenting conventional dialogue generation with an additional refinement stage enabling the dialogue system to correct potential hallucinations by querying the KG. NPH grounds dialogue generation

by constraining the flow of conversation to be supported by a valid path on the KG. To do so, the module combines a token-level hallucination critic that masks out entities of concern in an utterance, followed by a pre-trained non-autoregressive LM which prescribes contextual representations for each masked entity. This is then fed sequentially to an autoregressive LM to obtain output representations. These output representations can then be used to efficiently launch a query on the KG—effectively modelling dialogue as a signal being propagated on a local k-hop subgraph whereby locality is enforced through the conversation history—returning factually correct entities. Our proposed approach is applicable to any generated response whenever an available KG is provided and works without further fine-tuning. The high-level overview of our proposed approach is outlined in Fig.3.1 and exemplar machine-generated responses post-refinement are presented in Table 3.6. Our main contributions are summarized as follows:

- We conduct a comprehensive human study on hallucinations generated by state-of-the-art dialogue systems which reveals that the main mode of hallucinations is through the injection of erroneous entities in generated responses.
- We propose Neural Path Hunter, which leverages facts supplied by a KG to reduce hallucination in any machine-generated response.
- We empirically demonstrate that Neural Path Hunter substantially reduces hallucinations in KG-grounded dialogue systems with a relative improvement of 20.35% in FeQA, a QA-based faithfulness metric [32], and an improvement of 39.98% in human evaluation.

3.2 Hallucination in KG-grounded Dialogue Systems

We consider the task of generating factual and grounded dialogue when presented with auxiliary structured knowledge. In particular, we focus on factoids taken from multi-relational graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$, termed Knowledge Graphs (KG). Each KG consists

of a set of directed edge triples $t = \langle \texttt{[SBJ]}, \texttt{[PRE]}, \texttt{[OBJ]} \rangle$, where $\texttt{[SBJ]}, \texttt{[OBJ]} \in \mathcal{V}$ are nodes denoting subject and object entities and $\texttt{[PRE]} \in \mathcal{R}$ is a predicate that can be understood as a relation type. Broadly speaking, we say that a neural dialogue system is guilty of hallucinating whenever it generates a factual sentence that is not supported by a valid path in a k-hop subgraph $\mathcal{G}_c^k \subset \mathcal{G}$ of the original KG anchored around a context entity c.

As a starting point for our investigation, we study the various types of hallucinations a model may inject into an otherwise satisfactory response. Specifically, we explore the circumstances under which LMs are likely to exhibit unfaithful behaviour through misappropriation of entities (e.g. Barrack Obama was the President of Canada). Inspired by [106] for KG-grounded dialogue systems we hypothesize—among other possible mechanisms—hallucination can take form as either *intrinsic* or *extrinsic* to the provided KG.

Definition 3.2.1 (Extrinsic Hallucination)**.** *An extrinsic hallucination corresponds to an utterance that brings a new span of text that does not correspond to a valid triple in* $\mathcal{G}_c^k$.

From the perspective of definition 3.2.1, an utterance that might be partially faithful is still guilty of hallucination if there exists any injection of knowledge not authentically captured in $\mathcal{G}_c^k$. Despite this, external hallucinations can often be easier to identify due to their egregious nature. For example, the dialogue sample in Fig. 5.6 contains an external hallucination as the entity in question "Jay Roach" did not direct the movie "Titanic" and it is not supported within the 1-hop subgraph. On the other hand, the generated response may identify the correct set of entities but make false claims about their relationship which leads to the following definition.

Definition 3.2.2 (Intrinsic Hallucination)**.** *An intrinsic hallucination corresponds to an utterance that misuses either* `[SBJ]` *or* `[OBJ]` *in* $\mathcal{G}_c^k$ *such that there is no direct path between the two entities.*

Intrinsic hallucinations inject false information by condensing information from the KG in a wrong way. For instance, claiming that "Jay Roach" produced "Meet the Parents" is an incorrect association of the true relationship between these entities.

To ascertain the degree to which KG-grounded dialogue systems hallucinate and the nature of these hallucinations, we conduct a systematic evaluation by soliciting human judgement. We first fine-tune a LM on the OpenDialKG dataset [114] which contains a turn-based dialogue between two speakers on extracted triples from a known KG. The sequential nature of such turn-based dialogues grounded via extracted KG-triples effectively renders the entire conversation as a path traversed on the KG.

OpenDialkG OpenDialKG is a crowded-sourced English dialogue dataset where two workers are paired together to chat about a certain topic. The first speaker is asked to initiate the conversation about a given entity and the second speaker is tasked to form a factual response based a set of facts extracted from an existing KG, Freebase [7]. Those facts represent paths in the KG that are either 1-hop or 2-hop from the initial entity. Once the second speaker sends a response, the first speaker continues discussing the topic engagingly and new multi-hop facts from the KG are presented to the second speaker. The conversation can be regarded as traversing multiple paths in the KG. However, not all utterances within the same dialogue are grounded on facts from the KG. The second speaker can choose not to select a path from the KG to form an answer and instead forms a "chit-chat" response. Overall, the dataset consists of four domains: movie, music, sport and book where each second speaker's utterance is annotated with paths from the KG. The KG corresponds to a large subgraph extracted from Freebase with $\sim$ 1.2M triples (subject, predicate, object), $\sim$ 101k distinct entities and 1357 distinct relations. No official split is provided in the original dataset, and thus we randomly split the dataset in 80/10/10 for the train/valid/test, respectively. The data consists of 61778 train, 7933 valid and 7719 test. Some utterances in the dataset are chit-chat and thus are not annotated with a path from the KG. Thus, we

GPT2-KG	Hallucination			Faith.	Gen.
	Ex	In	B		
Greedy	**17.66**	**2.00**	**1.66**	**69.00**	9.66
Beam Search	18.33	3.33	4.00	68.00	6.33
Nucleus 0.9	25.33	4.00	2.33	64.66	**3.66**
Nucleus 0.5	23.33	5.33	4.33	59.90	7.00
Top20	28.33	7.00	5.00	55.00	4.66

Table 3.1: Human assessment of random 1500 GPT2 dialogue responses generated using OpenDialkg. "Ex", "In" and "B" mean extrinsic, intrinsic, and both hallucinations respectively. Each cell shows the mean percentage of responses with a specific dialogue property.

filter the dataset by keeping only the dialogue examples that are annotated with a path from the KG. We ended up with 23314 training examples, 2954 valid examples and 2954 test examples.

3.2.1 Modes of Hallucination

Experimental Protocol. As a demonstrative example, we use a pre-trained GPT-2 model [132] as the backbone of a neural dialogue system. To fine-tune GPT2, we concatenate the dialogue history, the KG-triples $\langle$`[SBJ]`, `[PRE]`, `[OBJ]`$\rangle$ and the ground truth response and then train the model to predict the next word in the response. To explore the effect of different decoding strategies and their impact in injecting hallucinations, we sample 300 responses from each decoding approach. We investigate greedy search, beam search, nucleus sampling [68] and top-k sampling [132] as representative decoding strategies.

For each dialogue sample, we crowd-source human judgement by soliciting evaluations from 3 different annotators from Appen[1], a high-quality annotation platform. Each annotator is tasked to first identify the presence of hallucination in the generated response when provided the dialogue history and KG triples. For samples where hallucination is present, we further ask the human annotators to identify whether the

[1]https://appen.com/

hallucination is extrinsic, intrinsic or both. If the response is not hallucinated, we ask them whether the response is faithful (i.e., supported by the triples) or generic (e.g., "I don't know about that"). The results of the human assessment are shown in Table 3.1. Overall, we report the average Krippendorf's alpha coefficient to be 0.72 on the annotator responses to the different questions which indicates high agreement. Using Table 3.1, we make the following key observations:

Remark 1. *Humans notice most hallucinations in KG-grounded dialogue systems are extrinsic.*

Remark 2. *A hallucination occurs the least in dialogue responses generated using a greedy decoding scheme. Conversely, top-k sampling results in the highest hallucination percentage (*40.33%*).*

Remark 3. *Increased diversity in response generation —i.e.(less generic), is positively correlated with an increase in hallucination e.g. Nucleus*=0.9*.*

Remark 1 indicates that the dominant mode of hallucination for all decoding strategies in KG-grounded dialogue systems is extrinsic rather than intrinsic. In fact, we find that in the OpenDialKG dataset, 54.80% of the responses contain extra entity mentions that are not supported by either $\mathcal{D}$ or $\mathcal{G}_c^1$ which may partially explain empirical observations. Remark 2 suggests that the model—when conditioned on factual knowledge—often assigns the highest probability mass to the correct response and sampling based on other distributions (e.g. top-k) invites hallucination in the generation process—a fact also observed in language modelling [78]. Remark 3 suggests an implicit trade-off between the different goals of response generation whereby improving the diversity of response can negatively impact its faithfulness. This reveals that in certain cases responses might be originally faithful to $\mathcal{G}_c^k$ but increasing diversity encourages the model to hallucinate. In light of these important observations, the main goal of this chapter is not necessarily to advance state-of-the-

art decoding methods but instead to instrument an efficient technique to identify hallucinations as well as retrieve the correct entities from the KG.

We seek to design a dialogue refinement system capable of fixing generated utterances such that they are semantically relevant given the conversation history *and* supported within a provided KG. To do so, we introduce Neural Path Hunter (NPH) a refinement strategy that can be easily applied to any generated response without retraining the model. NPH is composed of two modules: A token-level hallucination critic and an entity mention retriever. The first module flags and masks out hallucinated entities in an existing response and can be trained offline. The second module accepts masked representations identified by the critic and builds contextual representation of these problematic tokens which are then used to retrieve more faithful entities by running a query over $\mathcal{G}_c^k$. We assume the local k-hop subgraph is either provided or extracted based on the dialogue history. The following sections describe the data preparation, training, and inference procedures for these submodules.

3.2.2 Problem Formulation

Each instance in the dataset is composed of a dialogue history $\mathcal{D} = (x_1, \ldots, x_n)$, a set of j triples at turn n, $\mathcal{K}_n = (t_1, t_2, \ldots t_j)$ which together with $\mathcal{D}$ must be used towards generating the response $\bar{x}_{n+1}$. Here, each individual triple $t_i = \langle \texttt{[SBJ]}, \texttt{[PRE]}, \texttt{[OBJ]} \rangle$ is extracted from a provided KG. Thus, the task is to generate a response $\bar{x}_{n+1}$ that is faithful to a non-empty subset $M_n \subset \mathcal{K}_n$ —i.e., it can optionally talk about a few triples but not none. Specifically, the response $\bar{x}_{n+1}$ may contain entity mentions $m_i \in \mathcal{V}$ which indicates a factual response that potentially needs to be refined using NPH. For our purposes, it is most convenient to represent each mention as a tuple of three elements that indicates the beginning of the mention at position m_i^b and the end at position m_i^e. In other words, we represent an entity mention m_i as $m_i = (m_i, m_i^b, m_i^e)$. These entity mentions may not be faithful at all if they do not belong to either a `[SBJ]` or `[OBJ]` in M_n (extrinsic hallucination) or they could inject false relationships

between mentions via an unsupported path in $\mathcal{G}_c^k$ by incorrectly utilizing a [PRE] (intrinsic hallucination). We target and correct these unfaithful entities through retrieval over $\mathcal{G}_c^k$ in §3.3.1.

3.2.3 Token-level hallucination critic

To enforce faithfulness via refinement, we first identify the exact sources of hallucination in a given response. Based on the findings of human judgement in Table 3.1 and §3.2.1, we find hallucination errors in a dataset like OpenDialKG are often associated with entity mentions such as names of people, movies titles, locations, etc. To flag entities of concern, we design a token-level hallucination critic C that consumes $\mathcal{D}, \mathcal{K}_n, \bar{x}_{n+1}$ and outputs the set of hallucinated entity mentions M_c. To train C, we choose to cast the problem as a sequence labelling task where a binary label is predicted at each word position. As there is no labelled training data available for this task, we create a synthetic dataset consisting of ground truth dialogue samples and corrupted negative samples. We explore two corruption processes that convert a regular clean ground-truth response x_{n+1} to its corresponding hallucinated one $\hat{x}_{n+1}$ based on the type of hallucination we might expect to encounter —i.e. extrinsic and intrinsic.

1. **Extrinsic Negatives**. We replace each m_i in x_{n+1} with entities of the same type (e.g., person, location, etc...) but crucially not within $\mathcal{G}_c^k$ and the dialogue history $\mathcal{D}$.
2. **Intrinsic Negatives**. We simply swap every pair [SBJ] and [OBJ] in x_{n+1}. For example, the response "Crescendo was written by Becca Fitzpatrick" $\rightarrow$ "Becca Fitzpatrick was written by Crescendo" results in an intrinsic hallucination as in this case [PRE] is not bi-directional.

Overall, we apply a 60%/40% split of extrinsic versus intrinsic corruption strategies to the original train OpenDialKG to obtain a synthetic dataset to train C which is taken to be a pre-trained LM that is then fine-tuned on this binary classification task.

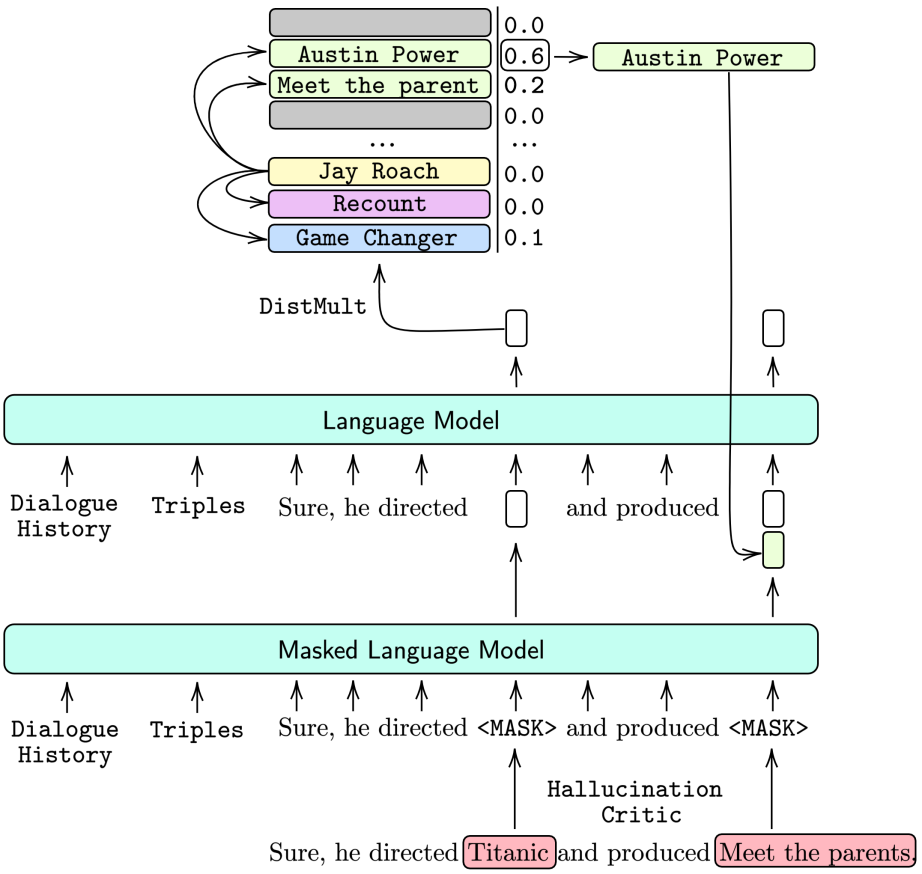


Figure 3.2: Entity Mention Retriever architecture.

3.3 Neural Path Hunter

3.3.1 Entity Mention Retriever

An overview of the Entity Mention Retriever is depicted in Fig. 3.2. Having identified entities of concern in $\bar{x}_{n+1}$, we now wish to craft a query that can be efficiently run over $\mathcal{G}_c^k$. To do so, we model the generated response $\bar{x}_{n+1}$ as a signal being propagated over $\mathcal{G}_c^k$ which serves to capture the highest probability paths starting from the context node c the conversation may take if it was faithful. The context node c is extracted from ground truth triples available in the dataset and or $\mathcal{D}$. In order to run an effective query over $\mathcal{G}_c^k$, it is critical that the representation of all flagged $m_i \in M_c$ and edge triples $\mathcal{E} \in \mathcal{G}_c^k$ are in the same representation space. Inspired by the Cloze task [163], we obtain contextual representations of all m_i's identified by the critic by

first masking them out before using a Masked Language Model (MLM). Operationally, we feed $\mathcal{D}$, $\mathcal{K}_n$, as well as the flagged set of entities to obtain contextual hidden state representations:

$$H = \text{MLM}(\mathcal{D}, \mathcal{K}_n, M_c) \tag{3.1}$$

As the MLM may return multiple hidden d-dimensional state representation for each $m_i \in M_c$, we simply apply a pooling operation to obtain a single representation for each entity —i.e. $h_i = \text{MaxPool}(h_b, h_e)$. To obtain the actual query q_i, we use an autoregressive LM which iteratively consumes an order dependent representation of h_i given by applying a learnable projection map $W : \mathbb{R}^{2d} \rightarrow \mathbb{R}^d$ to a concatenation of the current hidden state and the retrieved entity embedding e_{i-1} using previous query q_{i-1} as shown in Fig. 3.2,

$$q_i = \text{LM}(W(\text{concat}[e_{i-1}, h_i])),$$

KG-Entity Memory. Viewed another way, each q_i can be interpreted as a relation embedding for the masked position in $\bar{x}_{n+1}$. To effectively query $\mathcal{G}_c^k$, we must also represent all nodes in the same embedding space as q_i and in doing so effectively build a representation of $\mathcal{G}_c^k$ which we call KG-Entity Memory. We explore two approaches towards this goal. The first uses the final hidden layer of a pre-trained GPT2 to obtain initial embeddings for each node in $\mathcal{G}_c^k$ [2]. Our second approach uses CompGCN [169], which is a Graph Convolutional Network [82] purposely built for multi-relational data. We initialize the CompGCN network offline with GPT2 embeddings for all entities and relations in the full graph $\mathcal{G}$ before running a few rounds of message passing by optimizing for a standard relation prediction objective. Both approaches to KG-Entity memory embeddings can be further updated during training. Finally, to retrieve the correct entity for query q_i, we simply use a scoring function s to score every KG-Entity memory triple in $\mathcal{G}_c^k$ —i.e. $t_i = \langle c, q_i, \texttt{[OBJ]} \rangle$. The retrieved entity is the `[SUB]` or `[OBJ]` that achieves the highest score.

[2]Actually, GPT2 returns word piece representations and we use a pooling operation to get a single representation.

3.3.2 Training the Entity Mention Retriever

To train the Entity Mention Retriever, we augment the conventional maximum likelihood objective with an additional contrastive loss $\mathcal{L}_{\text{NCE}}$ that encourages faithful retrieval. In particular, we use Noise Contrastive Estimation (NCE) [63] which forces the Entity Mention Retriever to learn a scoring rule such that $s(t_i) > s(t'_i), \forall t_i \in \mathcal{E}, t'_i \in \bar{\mathcal{E}}$ where $t_i = \langle c, q_i, \texttt{[OBJ]} \rangle$ is the edge-triple based on KG-entity memory and $t'_i = \langle c, q_i, \texttt{[OBJ]}^- \rangle$ is a negative sample where $\texttt{[OBJ]}^-$ [3] is sampled from a corruption distribution over edge triples $\bar{\mathcal{E}}$ *not* in $\mathcal{G}_c^k$. To compute $\mathcal{L}_{\text{NCE}}$, we draw n negative samples uniformly over all entities for each query q_i.

$$\mathcal{L}_{\text{NCE}} = -\log\left(s(t)\right) - \log\left(s(t) + \sum_{j=1}^{n} s(t')\right).$$

At training time, we use teacher forcing [180]; first, we mask out all entity mentions within the gold response x_{n+1}, get their representations through a MLM and provide the ground truth entity mention concatenated with h_i at each time step in the LM. For the scoring function, we use DistMult [174] due to its simplicity in the absence of known structure over the modified triples e.g. translation, rotation, which are exploited in other popular scoring functions for KGs. By optimizing $\mathcal{L}_{\text{NCE}}$, we encourage the model to leverage the dialogue history, the position of the masked entity in x_{n+1}, and the k-hop subgraph to identify more faithful entities that are relevant to the conversation history. To train the Entity Mention Retriever, we thus jointly optimize $\mathcal{L}_{\text{NCE}}$ and $\mathcal{L}_{\text{MLE}}$ for the main language modelling task,

$$\mathcal{L} = \mathcal{L}_{\text{MLE}} + \lambda \mathcal{L}_{\text{NCE}}. \tag{3.2}$$

3.4 Experiments

We evaluate the ability of Neural Path Hunter towards reducing hallucinations in KG-grounded dialogue systems on the OpenDialKG dataset [114]. At present,

[3] or $\texttt{[SUB]}^-$ if c is an object

OpenDialKG is the only publicly available dataset that provides open-ended dialogue responses grounded on paths from a given KG, this is why we limit our experiments on this dataset. As there are no established metrics for this task, we consider a suite of task-specific and automated metrics to assess the different components of NPH and the degree of hallucination present. We use standard classification metrics such as F1-score, precision and recall to evaluate C and PPL to measure the quality of the LM. Similarly, we use retrieval metrics like Hits@k, Mean Rank (MR), and Mean Reciprocal Rank (MRR) to evaluate the Entity Mention Retriever.

3.4.1 Implementation Details

NPH: NPH is implemented using the Pytorch Huggingface Transformers library [184] and the Pytorch-lightning library [4]. Concretely, we use a small RoBERTa model [102] as the MLM and the base GPT2 model [132] as our autoregressive LM. During training, we use the Adam optimizer [79] with Dropout [157] on a batch size of 16 with a learning rate of 6.25×10^{-5} that is linearly decayed. The maximum dialogue history length is set to 3 utterances. The coefficient λ in Eq. 3.2 is set to 0.5. We varied the factor from 0.1 to 1 and 0.5 was chosen based on the best results on the validation set. The number of negative examples is set to 50 for SANS. The model early-stops at epoch 10 and we save the best model based on the validation set. Our hyperparameters search is done via greed search. The average runtime of this model is 4 hours.

Negative Candidates. We consider two different negative sampling strategies in order to compute $\mathcal{L}_{\text{NCE}}$: SANS [2] and In-batch-negatives. SANS selects hard negatives by leveraging the graph structure and selecting negative samples from a context entity's k-hop subgraph (e.g. $\mathcal{G}_c^1$). Meanwhile, In-batch-negatives considers the ground truth triple of each sample within a batch as a negative candidate for the other samples in the same batch. Using this approach, the number of candidates is equal to the batch

[4] https://github.com/Lightning-AI/lightning

size.

GPT2-KG: Similarly, we implement this baseline using the Pytorch Huggingface Transformers library and the Pytorch-lightning library. During training, we use the Adam optimizer [79] with Dropout [157] on a batch size of 32 with a learning rate of 6.25×10^{-5} that is linearly decayed. The maximum dialogue history length is set to 3 utterances. The model early-stops at epoch 6. The average runtime of this model is 2 hours.

AdapterBot and GPT2-KE: We use the code that's publicly available by the authors at https://github.com/HLTCHKUST/adapterbot and https://github.com/HLTCHKUST/ke-dialogue and we follow closely their training procedure described in [101] and [104]. We use the GPT2-KE with 9K iterations. The average runtime of these models is 3 hours.

Training for all models, including baselines, is done on an Nvidia V100 GPU 32GB and for inference, we use greedy search.

Hallucination Critic: We use a pre-trained RoBERTa-large classifier [102] provided by the Huggingface Transformers library [184]. The model was trained using the Adam optimizer with a learning rate of 2×10^{-5} for 5 epochs on one Nvidia V100 GPU 32GB. The average runtime of this model is 2 hours.

Hallucination Metrics. We consider 3 different hallucination metrics **M1-M3** that provide a multi-faceted measure of performance.

3.5 Hallucination Metrics

Although BLEU measures the extent to which the generated response is similar to the reference faithful response, it can be misleading in the case where the generated response is very distant from the ground-truth response but faithful to the knowledge

triples. We consider 2 other metrics that focus on measuring the degree of hallucination in the generated responses:

Hallucination Critic We use our trained token-level hallucination critic as a sentence-level hallucination detector. We consider the utterance as hallucinated if at least one token was identified as hallucinated. As input, the critic receives the dialogue history, the gold triples and the generated response and the output is a binary label indicating hallucination or not. To use this critic for the output of NPH, we augment the gold triples with the path extracted based on the Entity Mention Retriever.

FeQA Durmus *et al.* [32] has been shown successful in measuring faithfulness in the text summarization task. It generates questions from the candidate summaries and then answers them against the input documents. It measures the average F1 score against the gold answers from the document. Through asking and answering questions, FeQA measures the semantic correctness of the generated responses. To adapt FeQA to our dialogue task, we flatten each path into a pseudo sentence by joining the $\langle$`[SBJ]`, `[PRE]`, `[OBJ]`$\rangle$ with a simple space, e.g., [Crescendo, written by, Becca fitzpatrick] $\rightarrow$ "Crescendo written by Becca Fitzpatrick". We consider our document as the concatenation of $\mathcal{D}$ and all $\mathcal{G}_c^1$ triples and the candidate summary as the generated/refined response. FeQA takes a given generated grounded response as input, and generates questions. It then employs a QA system to answer the generated questions based on the knowledge the response was grounded in.

We use the code made publicly available by the authors [5]. A similar work to FeQA is QAGS [172] which corresponds to asking and answering questions to evaluate the factual consistency of summaries.

Negative Candidates. We consider two different negative sampling strategies in order to compute $\mathcal{L}_{\text{NCE}}$: SANS [2] and In-batch-negatives. SANS selects hard negatives

[5] https://github.com/esdurmus/feqa

by leveraging the graph structure and selecting negative samples from a context entity's k-hop subgraph (e.g. $\mathcal{G}_c^1$). Meanwhile, In-batch-negatives considers the ground truth triple of each sample within a batch as a negative candidate for the other samples in the same batch. Using this approach, the number of candidates is equal to the batch size.

3.5.1 Main Experimental Questions

Our experiments answer the following questions:

Q1) Identifying Hallucinations. Can C identify both extrinsic and intrinsic hallucinations?

Q2) Reducing Hallucinations. Is NPH effective in reducing hallucinations?

Q3) Query Generation. Can NPH retrieve the correct entities and is $\mathcal{L}_{\text{NCE}}$ important to learn query representations q_i?

Q4) Impact of MLM and Critic. Is MLM essential to our training strategy or can we only use an autoregressive LM? Analagously, can we simply bypass the critic during refinement?

Q5) Impact of global graph structure. Is the global graph structure important for learning KG-Entity memory representations?

3.5.2 Results

Throughout our experiments, we rely on three representative baselines for response generation: GPT2-KG, AdapterBot [101], and GPT2-KE [104]. GPT2-KG is a small pre-trained GPT2 model [132] fine-tuned on the dialogue corpus. AdapterBot uses a fixed backbone conversational model such as DialGPT [199] and encodes multiple dialogue skills via different adapters [71]. Both GPT2-KG and AdapterBot process inputs by concatenating $\mathcal{D}$, $\mathcal{K}_n$ and the generated response. GPT2-KE on the other hand uses a GPT2 model trained on a knowledge-augmented training set.

Q1: Identifying Hallucinations

Analogous to the study conducted in §3.2.1, we ask humans to identify the span of text that is hallucinated w.r.t. to the given triples in 500 responses generated greedily from GPT2-KG. We report the average Krippendorf's alpha coefficient to be 0.73 on the annotator responses. Table 3.2 outlines our results. To explore the robustness of our corruption strategies, we fine-tune a large RoBERTa model [102] on three different synthetic datasets: (i) `RoBERTa-Extrin` corresponds to the negative examples crafted using an extrinsic hallucinations, where entity mentions are first extracted using the SpaCy NER tagger [69]. (ii) `RoBERTa-Intrin` consists of negative examples that contain intrinsic hallucinations. (iii) Finally, `RoBERTa-Intrin-Extrin` corresponds to examples that were either corrupted using an extrinsic or intrinsic strategy but not both simultaneously. For (i) and (ii), the examples are obtained by corrupting the full train OpenDialKG data. We observe that `RoBERTa-Intrin-Extrin` achieves the highest F1 (70.35%), compared to the classifiers trained on the first two synthetic datasets. Such a result highlights that our `RoBERTa-Intrin-Extrin` classifier can indeed detect both kinds of hallucinations and also that our corruption strategies are effective. In the rest of the experiments, we take `RoBERTa-Intrin-Extrin` as the hallucination classifier C.

Model	**Precision**	**Recall**	**F1**
RoBERTa-Intrin	44.9	32.54	37.73
RoBERTa-Extrin	68.65	46.94	55.76
RoBERTa-Intrin-Extrin	**83.05***	**61.02***	**70.35***

Table 3.2: Performance of the hallucination critic on the 500 human-annotated data (* p-value < 0.001)

Q2: Reducing Hallucinations

We evaluate the ability of NPH in fixing hallucination in generated responses in the three response generation baselines. We also perform ablation for each model

using the different components of NPH. We present the results in Table 3.4 which show the degree of hallucination prior to and after applying NPH on each response generation method. We find that NPH consistently performs favourably in reducing hallucination across FeQA and the hallucination Critic. In particular, we observe that the strongest iteration of each baseline model is the original model paired with the full NPH module. For example, in AdapterBot, NPH decreases the Critic score by 8.17 points and increases faithfulness by 6.67 points on FeQA. With respect to BLEU scores, we observe inconsistent performance across the different baselines with AdapterBot+NPH incurring a marginally higher score. While we use BLEU as a proxy for faithfulness, it is still an imperfect measure as it is computed solely between the n-gram overlap between a reference and generated text which neglects the important fact that there is a multitude of different ways to generate a faithful response w.r.t. a KG.

	Model	Neg. candidates	PPL	Hits@1	Hits@3	Hits@10	MR	MRR
GPT2-Emb	NPH	SANS	**8.56**	**0.73**	**0.92**	**0.99**	**1.76**	**0.83**
		In-Batch Negatives	8.67	0.42	0.75	0.94	3.08	0.68
	NPH-w/o nce	-	9.64	0.02	0.05	0.1	35.49	0.07
	NPH-w/o mlm	SANS	9.73	0.47	0.76	0.96	2.83	0.64
		In-Batch Negatives	9.70	0.20	0.43	0.75	9.22	0.36
CompGCN-Emb	NPH	SANS	**8.99**	**0.13**	**0.26**	**0.52**	**14.27**	**0.25**
		In-Batch Negatives	10.04	0.08	0.17	0.43	15.75	0.16
	NPH-w/o nce	-	10.61	0.04	0.12	0.27	26.50	0.12
	NPH-w/o mlm	SANS	9.63	0.08	0.21	0.47	15.52	0.20
		In-Batch Negatives	9.64	0.02	0.05	0.16	80.52	0.07

Table 3.3: Ablation studies on Neural Path Hunter on the gold responses from the OpenDialKG test data.

Q3: Query Generation

We now investigate NPH's ability to retrieve the correct entity using the crafted query. We present the results in Table 3.3 along with different ablation studies. We find that key metrics such as Hits@3 and Hits@10 are nearly saturated when using the complete

NPH module with GPT2 embeddings for the KG-Entity memory. Furthermore, we notice that all retrieval metrics drop dramatically (e.g.↓ 70 Hits@1) when $\mathcal{L}_{\text{NCE}}$ is omitted. Finally, we observe that SANS negatives lead to lower perplexity and better retrieval performance across the board. This is unsurprising since local negative samples are known to be harder and thus provides a richer learning signal [2].

Model	**FeQA ↑**	**Critic ↓**	**BLEU**
GPT2-KG	26.54	19.04	**11.79***
+ NPH	**28.98***	**11.72***	11.29
+ NPH-w/o nce	26.02	17.91	10.98
+ NPH-w. CompGCN	26.89	15.41	11.10
+ NPH-w/o MLM	27.01	15.02	10.88
+ NPH-w/o critic	18.23	19.65	6.49
AdapterBot	23.11	26.68	10.56
+ NPH	**27.21***	**18.51***	**10.74***
+ NPH-w/o nce	24.02	25.02	9.98
+ NPH-w. CompGCN	25.83	20.23	10.11
+ NPH-w/o MLM	26.02	21.04	10.06
+ NPH-w/o critic	16.21	27.22	5.64
GPT2-KE	19.54	28.87	**6.24***
+ NPH	**26.21***	**20.34***	6.06
+ NPH-w/o nce	20.34	.32	5.89
+ NPH-w. CompGCN	23.23	21.21	6.01
+ NPH-w/o MLM	24.01	22.40	5.99
+ NPH-w/o critic	15.89	30.71	3.49
Gold response	33.34	5.2	-

Table 3.4: Measuring the degree of hallucination of different models pre and post-refinement on generated samples based on the OpenDialkg test data. A higher FeQA score indicates an increase in faithfulness. The hallucination Critic (Critic) measures the percentage of hallucinated responses in the dataset. (* p-value < 0.001). NPH uses GPT2 emb. for the KG-Entity Memory.

Q4: Impact of MLM and Critic

We now gauge the importance of using MLM and Critic within NPH. To assess the MLM component, we replace each contextual representation $m_i \in M_c$ with randomly initialized values. We highlight our findings in Table 3.4 where NPH-w/o MLM performs worse than NPH across all models. Investigating further in Table 3.3, we observe that performance without MLM degrades substantially (e.g. $\downarrow$ 26 Hits@1) when using pre-trained GPT2 embeddings as entity memory and similarly for CompGCN embeddings. These findings suggest that MLM facilitates the learning of rich masked representations that are useful in downstream applications, a fact which is in line with other works that leverage MLM [141, 27, 74]. To judge the impact of the critic, we mask out all entity mentions as opposed to only masking out potential hallucinated ones during refinement. In Table 3.4, we find that NPH-w/o CRITIC performs the worst in every metric compared to all baselines which underlines that simply masking all entities—hallucinated or otherwise—in a response is not a productive strategy for effective refinement.

Q5: Impact of global graph structure

We now investigate the representation of entities in our KG-Entity Memory. We explore two variants: 1) Initializing embeddings as the output of a pre-trained GPT-2 model. 2) Utilizing node embeddings learned by a CompGCN network trained on a standard relation prediction task over the entire graph $\mathcal{G}$. In both these approaches, the embeddings are updated throughout training using Eq. 3.2. As per Table 3.3, we notice a dramatic difference in both perplexity and retrieval performance in favour of using simply the output of a pre-trained GPT-2 model. Such a result may be reconciled by noticing that any specific turn in dialogue local information (e.g. previous turn)—as conversation topics may drift—is significantly more important to generate a faithful response. Thus, enriching entity embeddings with global structure in $\mathcal{G}$ is less beneficial than aligning $\mathcal{G}_c^k$ with the representation space of the autoregressive LM,

which for us is also GPT2.

Model	Hallucination	Fluency
GPT2-KG	97.5 ± 0.6	92.5 ± 1.6
GPT2-KG (+ NPH)	**56.5** ± 1.2	88.5 ± 0.7
AdapterBot	95.5 ± 0.8	90.5 ± 0.4
AdapterBot (+ NPH)	**59.0** ± 0.5	87.5 ± 1.2
GPT2+KE	97.0 ± 0.2	91.5 ± 0.7
GPT2+KE (+ NPH)	**58.5** ± 0.6	86.0 ± 0.9

Table 3.5: Human Evaluation on 1200 responses (200 × 6) from different response generation baselines.

3.5.3 Human Evaluation

In addition to the automated hallucination metrics, we conduct human evaluation to assess NPH's ability to reduce hallucination. We provide human annotators with 200 hallucinated responses per baseline (§3.5.2) as identified by our hallucination critic §5.6.1. The faithfulness of each response is evaluated by 3 humans who are provided $\mathcal{D}$, $\mathcal{K}_n$, and the retrieved path from $\mathcal{G}_c^k$. We further request annotators to evaluate the fluency of the responses before and after refinement. Results are depicted in Table 3.5. We see that the hallucination critic achieves a precision of 97.5% for GPT2-KB responses, 95.5% for AdapterBot and 97.0% for GPT2-KE. In contrast, generation methods when paired with NPH reduce hallucinations by a large margin 42.05% for GPT2-KB responses with a marginal drop in fluency (4.32%). We also observe similar performance gains for responses generated from AdapterBot and GPT2-KE.

Human Evaluation of NPH responses Analogous to evaluating modes of hallucination, we solicit human evaluation from Appen[6] where we train English-speaking

[6]https://appen.com/

annotators for the task before starting the evaluation process. To evaluate the responses generated by our response generation baselines, annotators were presented with $\mathcal{D}$, $\mathcal{K}_n$ and the generated response. And, to evaluate NPH's responses, annotators were presented with $\mathcal{D}$, $\mathcal{K}_n$, the retrieved path from $\mathcal{G}_c^k$ and the refined response. Humans were asked to answer the following questions:

1. Is this response hallucinated with respect to $\mathcal{K}_n$? (Most definitely, Not at all)
2. Is this a fluent response, i.e., a response that's grammatically correct? (Most definitely, Not at all)

In total, humans evaluated 1200 responses: 600 responses (200 from each response generation baseline before refinement) and 600 responses after refinement.

3.6 Error Analysis

To gain insight into the potential shortcomings of NEURAL PATH HUNTER, we conduct an error analysis on refined responses that still contain undesirable hallucinations. Examples of failed refinements using NPH are listed below. Recall that for effective retrieval NPH requires oracle access to $\mathcal{G}_c^k$ which pre-supposes the existence of the correct entity in the subgraph. However, based on the examples below, we observe that many of the failed retrievals correspond to entities that might exist in $\mathcal{G}$ but are critically not supported within $\mathcal{G}_c^k$. To highlight this point, let us consider the following example:

Previous turn: Could you recommend a book similar to Thirteen Reasons Why?

Gold triple:[['Thirteen Reasons Why', 'has genre', 'Young-adult fiction']

GPT2-KB Response: Sure, there is a book called **The Sea of Monsters** by John Green.

Critic: John Green.

Context nodes: Thirteen Reasons Why, Young-adult fiction

Top-1 Path:[[Thirteen Reasons Why, written by, Jay Asher]

NPH response: Sure, there is a book called The Sea of Monsters written by Jay Asher.

In this example, we notice that the original response contains hallucination as "John Green" did not write the book titled "The Sea of Monsters". Moreover, to fix this response NPH has access to $\mathcal{G}_c^1$ which contain the context nodes "Thirteen Reasons Why" and "Young-adult fiction" but not the entity of interest "The Sea of Monster". As a result the proposed fix using NPH is the wrong entity "Jay Asher". It is important to note that the perceived mistakes by NPH is in fact a limitation of our assumption that the correct entities not only exist in $\mathcal{G}$ but can be extracted by running a query on the extracted k-hop subgraph. If this fails to be the case in practice, then we cannot hope for NPH to retrieve the correct entity. We hypothesize that extending NPH to reason over multiple subgraphs or even $\mathcal{G}$ in an efficient manner will serve to alleviate this failure mode.

In addition to using the incorrect subgraph for retrieval, another source of errors is caused by the failure to identify the hallucinated entities to fix in the first place. This can occur in practice as the Critic C is not perfect and may occasionally miss hallucinated entities and thus propagate the error forward. The example below illustrates this failure mode:

Previous turn: Yes, I watched them. Could you tell me more about The Heart Is Deceitful Above All Things, please.

Gold triple:[The Heart Is Deceitful Above All Things, starred actors, Jimmy Bennett]

GPT2-KB Response: Sure, it was written by Tony Kaye and stars Scarlett Johansson.

Critic: Scarlett Johansson.

Context nodes: The Heart Is Deceitful Above All Things, Jimmy Bennett

Top-1 Path:[The Heart Is Deceitful Above All Things, starred actors, Jimmy Bennett]

NPH response: Sure, it was written by Tony Kaye and stars Jimmy Bennett.

We believe a stronger Critic or other mechanisms to identify token-level hallucinations can improve the downstream performance of NPH.

The final source for error—while not frequent—is the fluency of the refined response. We find that in an effort to remove hallucinations, other undesirable properties such as a lack of grammaticality may be introduced. The example below demonstrates such a failure mode. This error can be mitigated if we associate each node in the graph by an entity type.

Previous turn: Sure. Isn't he married to Coleen Rooney

Gold triple:[[United Kingdom, Country of nationality, Wayne Rooney]

GPT2-KB Response: No, Yes, he is married to Caroline Wozniacki.

Critic: Scarlett Johansson.

Context nodes: United Kingdom, Wayne Rooney

Top-1 Path:[United Kingdom, Country of nationality, Wayne Rooney

NPH response: Yes, he is married to United Kingdom.

3.7 Related Work

Knowledge Graphs. Building large-scale repositories of knowledge has been one of the principle directions of research in artificial intelligence since the inception of the field [119, 118]. Often represented as large scale multi-relational graphs, KGs have seen wide application in a variety of domains, such as question answering [187, 65], and natural language processing [10, 190] to name a few. Beyond academic research, public KG's like FreeBase [12] have been invaluable in industrial applications forming symbolic backbones of most important products [151]. KG's have also risen in prominence in the context of dialogue models that propose to explicitly embed symbolic knowledge representations into a neural embedding space [102, 202, 114, 200, 185]. Niu *et al.* [121] use a knowledge retriever component that conditions the response by retrieving relevant facts from the KG based on the current utterance. Similarly, Young *et al.* [189] and Zhou *et al.* [200] use a commonsense KG to inject commonsense knowledge into the response of the conversational model. Tuan *et al.* [168] explore the effects of using a dynamic KG in the dialogue model. On the other

hand, Moon *et al.* [114] propose a conversational reasoning model that traverses a large scale KG to retrieve a relevant path given a starting node and a classifier to predict the next node a response show follow. Unlike the KG path traversal problem, this work focuses on removing hallucinations in generated responses using a KG.

Hallucination. The injection of false information is a well-known phenomena in data-to-text generation [165, 29, 126], machine translation [85, 91], image captioning [142], machine summarization [106, 32] and question answering [49]. In the context of dialogue systems, Dušek *et al.* [34, 35] demonstrate that state-of-the-art natural language generation (NLG) models can hallucinate by missing important entities. Few NLG models have been proposed to cope with the issue, but are often custom-made for task-oriented dialogue [6]. Recently, little progress has been made for studying hallucination in open-domain dialog systems. Dziri *et al.* [42] study hallucination in knowledge-grounded dialogue systems and introduce a the BEGIN benchmark for measuring groundedness in dialogue systems. Finally, Rashkin *et al.* [135] propose a dialogue system that is more faithful to the source knowledge by adding control tokens at training time that guide the model towards generating more objective sentences which have higher overlap with the source.

3.8 Conclusions

In this chapter, we investigate the open problem of hallucination in KG-grounded dialogue systems and demonstrate that these models are more susceptible to extrinsic hallucinations which predominantly manifest as the injection of erroneous entities. To tackle this challenging problem, we propose a new module Neural Path Hunter that aims to enforce faithfulness in KG-grounded dialogue systems by identifying and refining hallucinations via queries over a k-hop subgraph. We empirically observe that NPH is capable of reducing hallucination when paired with a number of base dialogue models with relative improvements of 20.35% over vanilla GPT2 on FeQA. Our findings also reveal the crucial role the representation of the local subgraph plays

History	A:	I love Shakespeare. Romeo and Juliet is my favorite.
	B:	I like that one as well. The play was written by Kylie Scott.
	A:	She also wrote Deep, did she not?
GPT2-KG	B_{gen}	Yes she did. She also wrote The Tragedy of King Richard the Third.
Gold knowledge	T_1	[Deep, written by, Kylie Scot]
Top-5 Paths	T'_1:	[Play, written by, Kylie Scott]
	T'_2:	[Dirty, written by, Kylie Scott]
	T'_3:	[Lick, written by, Kylie Scott]
	T'_4:	[Deep, written by, Kylie Scott]
	T'_5:	[Trust, written by, Kylie Scott]
NPH response	B_{fix}	Yes she did. She also wrote Play.
History	A:	Do you know of a movie directed by Jean Pierre Jeunet?
GPT2-KG	B_{gen}	Jean-Pierre jeunet directed The Bourne Ultimatum.
Gold knowledge	T_1:	[Micmacs, written by, Jean Pierre Jeunet]
	T_2:	[Micmacs, has genre, French]
Top-5 Paths	T'_1:	[Alien: Resurrection, directed by, Jean Pierre Jeunet]
	T'_2:	[A Very Long Engagement, directed by, Jean Pierre Jeunet]
	T'_3:	[Amélie, directed by, Jean Pierre Jeunet]
	T'_4:	[The Extraordinary Voyage, starred actors, Jean Pierre Jeunet]
	T'_5:	[Micmacs, written by, Jean Pierre Jeunet]
NPH response	B_{fix}	Jean-Pierre jeunet directed Alien: Resurrection.

Table 3.6: Selected responses based on GPT2-KG test responses before and after applying Neural Path Hunter. The span of texts highlighted in red indicate the hallucinated entity mentions whereas the ones highlighted in green indicate the retrieved correct entity mentions.

as external memory compared to the full global graph. In this work, we consider a paired KG aligned with dialogue but in many other applications, such dialogue to KG alignment may be difficult to easily obtain necessitating the usage of the full graph which is interesting direction for future work.

Chapter 4

Evaluating Attribution in Dialogue Systems

Neural language models [9, 170, 132, *inter alia*] often form the backbone of open-ended dialogue systems [183, 197, 143, 1]. Utterances sampled from such language models sound natural, as reflected in these systems' high scores in human evaluations focused on measures such as "engagingness" or "human-likeness" [146]. While fluent, however, the responses generated by these systems often contain statements that are hallucinated ([166, 106, 39, 150]; see Figure 4.1 for an example). Progress towards models that do not exhibit this issue requires evaluation metrics that can quantify its prevalence. In this chapter, we introduce a benchmark that can be used to assess attribution in knowledge-based dialog systems; following Rashkin *et al.* [136], we define an attributable response[1] as one connected to textual evidence that supports the entirety of the response.

4.1 Introduction

A number of modelling approaches have recently been proposed to increase attribution in knowledge-grounded dialog systems [135, 150, 39, 43]. Progress in this area crucially relies on metrics that can measure the attribution of the text generated by the system; and indeed, recent work has developed automated metrics with relatively high

[1]Attribution is sometimes referred to as faithfulness [17, 32, *inter alia*].

correlations with human annotations, potentially paving the way for alternatives to expensive human evaluations [70, 39, 43]. Yet our understanding of these recently proposed metrics, as well as more established ones, remains limited, for two reasons. First, comparisons between automated metrics and human judgments rely on small-scale datasets with a few hundred examples. This results in high variance in our estimate of the correlation coefficient and a limited ability to measure performance on infrequent example types [54]. Second, the correlation with human scores does

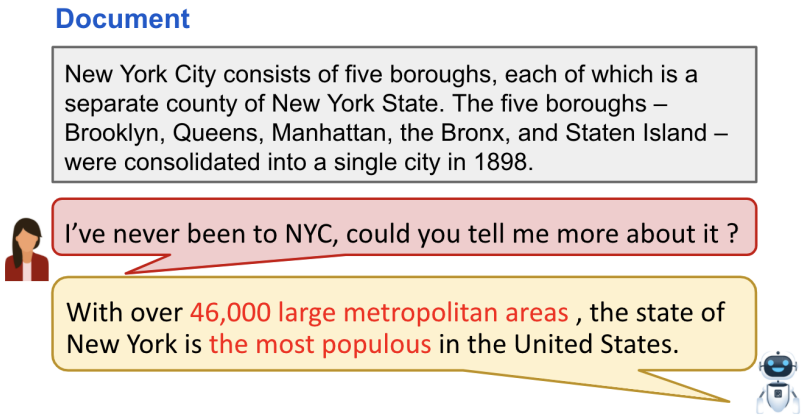


Figure 4.1: An example of a response generated by the GPT2 language model fine-tuned on the Wizard of Wikipedia dataset [30]. The phrases in red are "hallucinations" unsupported by the background document.

not sufficiently determine the efficacy and robustness of automatic metrics produced by neural networks: such learned metrics—like other properties learned by neural networks—can be susceptible to spurious correlations that fail to generalize to more challenging cases. To address these limitations, we introduce a large-scale resource, the Benchmark for Evaluation of Grounded INteraction (BEGIN), for meta-evaluation of metrics designed to evaluate grounded dialogue. In other words, the goal of this benchmark is to determine to what extent current evaluation metrics fulfill their purpose.

We define a taxonomy dividing knowledge-grounded dialogue responses into three broad categories—*fully attributable*, *not fully attributable*, and *generic*—and ask humans to classify a large set of utterances produced by dialogue systems with

this taxonomy. The motivation for the *generic* category we introduce—which is assigned to utterances such as "*Sorry, I'm not sure about this topic*"—is the intuition that evaluation metrics should not treat the basic elements of a natural-sounding conversation, such as backchanneling or acknowledgment [59, 158, 15], as equally undesirable as a misleading unattributable statement. In real-world scenarios, it is preferable for a model to acknowledge its ignorance instead of producing hallucinated content which may lead to the spread of disinformation.

Using this taxonomy, we then collect high-quality human annotations for 12k examples generated by four language-model-based dialogue systems, each trained on three different knowledge-grounded dialogue corpora. Examples of machine-generated responses along with labels are presented in Table 4.1. We use this benchmark to evaluate multiple existing automatic metrics including word-overlap measures, embedding-based measures, metrics based on Question Answering (QA) systems, and ones based on Natural Language Inference (NLI). We also propose a classifier trained on an adversarially generated dataset we create. We find that all metrics inadequately measure attribution and all rely on spurious correlations to a large extent. In particular, the metrics tend to misidentify cases that are attributable but highly abstractive, as well as cases that are not fully attributable but use multiple words from the evidence document (i.e., unattributable but extractive). We also find that the metrics fail to measure attribution under distribution shift, scoring responses that pertain to relatively long knowledge sources the lowest. These results are in line with the robustness issues reported for other natural language generation metrics, despite the high correlation of those metrics with human judgments [33, 54, 52, 188]. We hope that BEGIN will facilitate progress toward more robust metrics for grounded dialogue response generation.

	Category	Example	
WoW	Fully Attributable	Knowledge:	**Raised in Compton, California, Lamar embarked on his musical career as a teenager** under the stage name k-dot, releasing a mix-tape that garnered local attention and led to his signing with indie record label top dawg entertainment.
		History:	oh, yeah. I believe I've heard about that. The album has to be good then for sure. Do you know anything else about him or rap?
		T5 Response:	Oh yes, I know that the singer was raised in Compton, California and started on his musical career as a teenager.
CMU-DoG	Generic	Knowledge:	Throughout the film, the narrative is intercut with scenes from depositions taken in the Winklevoss twins' and Saverin's respective lawsuits against Zuckerberg and Facebook. ... After everyone leaves, Zuckerberg is shown sending a friend request to Albright on Facebook ...
		History:	The ending seems cools. Like after all of that happens he just goes to hang out with his kids.
		GPT2 Response:	What else do you know about Zuckerberg?
TopicalChat	Not Attributable	Knowledge:	Television (TV) is a telecommunication medium used for transmitting moving images in (black and white) .... Ringo Starr, George Carlin, and Alec Baldwin have all narrated Thomas the Tank Engine for at least 52 episodes each. There are over 750 episodes of "Pokémon"
		History:	no, I haven't watched Pokemon. My sons use to play with when they were little. Now, they use their phones looking for Pokemon in the neighborhood. It is funny.
		DoHA Response:	Pokemon appeared in TV in 1952 and it was created by Ringo Starr and directed by George Carlin. In total, there were 100 episodes.

Table 4.1: Examples of each of the three categories of responses included in BEGIN. For each category, we provide an example drawn from one of the four models trained on one of the three corpora (of course, all 12 models generated all three types of responses). The dialogue corpus used to train the model and generate the response is listed vertically. Text highlighted in green indicates information that is attributable to the knowledge; text in blue does not convey any information; and text in red is hallucinated and cannot be attributed to the knowledge.

4.2 Task, Datasets and Models

In knowledge-grounded response generation, the system is given a dialogue history $\mathcal{H} = (u_1, \ldots, u_{n-1})$, and knowledge $\mathcal{K}_n = (k_1, \ldots, k_j)$ at turn n, and is expected to generate a response $\bar{u}_n$ that is coherent with $\mathcal{H}$ and attributable to a non-empty subset $M_n \subset \mathcal{K}_n$. Similar to the conversational QA task [22, 138], the system is expected to use knowledge to respond to the user query. However, since the previous utterance may be an open-ended statement rather than a direct question (see the second and third examples in Table 4.1), there is a wider range of possible types of informative replies compared to the conversational QA task.

BEGIN consists of responses generated by language-model-based systems trained to perform this task. This section describes the models we train on this task and the corpora we use to train them.

4.2.1 Dialogue Datasets

For all three datasets, we use the training portion to train the model, the development set to tune hyperparameters, and the test set to generate the responses that are then annotated and included in the final BEGIN benchmark. We used Wizad of Wikipedia [30], CMU-DoG [201], and TopicalChat [57].

4.2.2 Dialogue Models

We consider the outputs of four different dialogue systems; by selecting a relatively wide range of systems, we hope to encounter a range of attribution errors. Two of the systems are based on plain language models, GPT2-base [132] and T5-base [133]. The remaining two systems, DoHA [129] and CTRL-DIALOG [135], are specifically designed as knowledge-grounded dialogue systems. DoHA augments a BART-based conversational model [92] with a two-view attention mechanism that handles the encoded document and the dilaogue history separately during generation. CTRL-DIALOG augments T5-base with control tokens [78] that guide the generation towards

less subjective and more grounded content. We trained these models to generate responses based on a concatenation of two inputs: an evidence span (the knowledge snippet) and the dialogue history (we only use the previous turn u_{n-1}).

4.3 Annotations

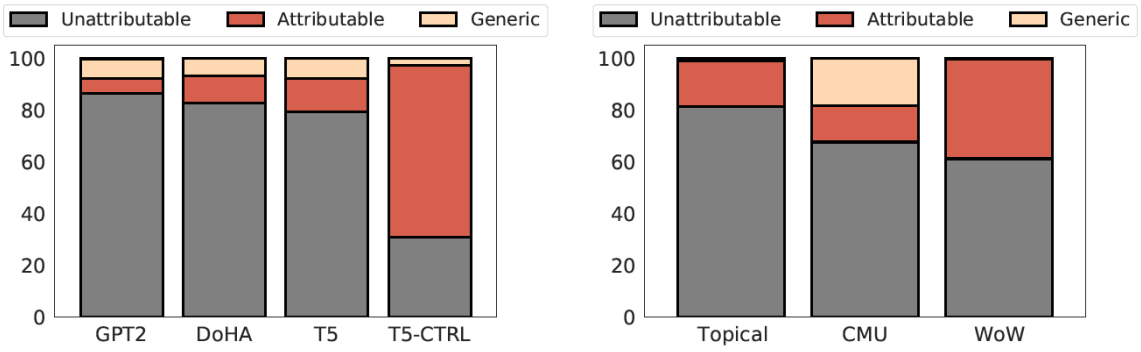
We next describe the human annotations we collected for the utterances generated by the models described in Section 4.2.

4.3.1 Taxonomy of Response Types

We classify responses into three broad categories:

Fully Attributable These are responses that convey information that can be completely supported by the provided document; this property has been referred in the literature to as faithfulness [135, 106, 39, 32] and attribution [136]. In our annotation set-up, we use similar definitions to the Attributable to Identifiable Source (AIS) framework of Rashkin *et al.* [136]. The full framework in that paper consists of a two-stage annotation process in which annotators first filter out responses that are deemed to be too vague or ill-formed to be evaluated for attribution. Since Rashkin *et al.* [136] found that more than 90% of the conversational responses in their study were interpretable, we have our annotators focus solely on attribution.

Not Attributable These are responses that contain at least some information that cannot be verified given the evidence, regardless of whether that information is factually true in the real world. This includes statements that are relevant but not fully supported by the background information (hallucinations), statements that explicitly contradict the background information, and off-topic responses about information completely external to the evidence sources. In a pilot study we attempted to separate these three subcategories, but the boundaries between them turned out to be difficult



(a) Breakdown by model class.

(b) Breakdown by dialogue corpus used for training.

Figure 4.2: Breakdown of BEGIN response categories across models (left) and training corpora (right).

to define and annotate.

Generic Responses that fall into this category are general enough to fit into a large number of possible contexts [95]. Examples include "I don't know about that" and "Hello there!". Even when the responses are ostensibly about the same topic as the document, they are vague and do not provide new information. Nevertheless, such responses may be useful for various conversational purposes: back-channeling, expressing uncertainty, or diverting the conversation from ambiguous or controversial topics.

4.3.2 Collecting Prompt-Query-Reply Triples

As described in Section 4.2, we collect data using outputs from four models—T5, GPT2, DoHA, and CTRL-DIALOG. We train a version of each model on each of the three datasets (WOW, TOPICALCHAT and CMU-DOG) and generate responses using the test portion of the dataset. For more details on training and hyperparameters, refer to Appendix 4.6. We select at least 1000 examples from each dataset-model pair. We filter and remove toxic responses using the Google Perspective API. This yields 12288 examples in total.

4.3.3 Annotating Prompt-Query-Reply Triples

We present annotators with a knowledge snippet $\mathcal{K}$, the previous turn u_{n-1} and a generated response $\bar{u}_n$, and ask them to select which of the three categories fits $\bar{u}_n$ best. To obtain high quality data, we assign three annotators to each example and report results based on majority vote. We exclude examples where each of the three annotators assigned a different category, making it impossible to compute a majority vote.

Begin Annotation Protocol Each worker was given a document, previous turn in a conversation and a generated response (either by T5, GPT2, DoHA or CTRL-Dialog). They were asked to evaluate the response as either fully attributable, not attributable, or too generic to be informative. They also were provided with multiple examples with explanations for each category. The exact instructions were as follows:

> Which of these best describes the highlighted utterance?
>
> - Generic: This utterance is uninformative (too bland or not specific enough to be sharing any new information)
> - Contains *any* unsupported Information: This utterance is sharing information that cannot be fully verified by the document. It may include false information, unverifiable information, and personal stories/opinions.
> - *All* information is *fully* supported by the document: This utterance contains only information that is fully supported by the document.

Annotation Quality To ensure that the annotators understood the task, we use the following manual quality control procedure. In the first stage, we train the annotators by running two pilot annotation batches ($\sim$ 100 examples each). After each batch, we manually grade the answers for compliance with instructions, and provide feedback explaining any misconceptions. After the training stage, we launch the main annotation round for the full set of 12k examples. During this round, we intermittently check responses after every 3k completed annotations to examine the

annotation quality. This procedure resulted in high inter-annotator agreement (a Krippendorff's alpha of 0.7).

4.3.4 Dataset Analysis

BEGIN is intended as a test benchmark; as such, it does not have a training portion: We only create development (10%) and test (90%) partitions. We include examples from BEGIN in Table 4.1 along with the label breakdown. Overall, the models generated a substantial number of unattributable responses (70%). As Figure 4.2 (right) shows, this proportion was higher for GPT2, DOHA, and T5, whereas CTRL-DIALOG generated the lowest proportion of unattributable responses (30.8%). This indicates that CTRL-DIALOG, which is explicitly designed to discourage unattributable responses, is moderately successful at its goal. Figure 4.2 (left), which breaks the results down by training corpus, shows that models trained on TOPICALCHAT produce the highest amount of unattributable responses followed by CMU-DOG and WOW. This is consistent with recent analyses on WOW, CMU-DOG and TOPICALCHAT which revealed that more than 60% of the ground-truth responses are unattributable to the knowledge [41, 136].

4.3.5 The Need to Measure Attribution

Our analysis of the responses produced by the systems we trained highlights the potential pitfalls of language-model-based dialogue systems, especially when deployed in real-world scenarios across a broad range of domains where hallucinations pertaining to vital information may produce undesirable user experiences—e.g., healthcare [90, 75] and education [186, 84]—and underscores the need for progress on both the modeling and the evaluation side. Neural dialogue systems are optimized to mimic the distributional properties of the human-generated dialogue corpus used to train them. Because humans often include unattributable information in their utterances, language models trained on those corpora can replicate and perhaps even amplify the

prevalence of unattributable responses at test time [77, 41]. These findings call for robust evaluation metrics to uncover actionable insights about best practices of using such models and benchmarks. We hope that BEGIN will, as an evaluation benchmark, promote a strict standard for evaluation metrics, laying the ground for trustworthy dialogue systems.

4.4 Evaluating Evaluation Metrics

We next use BEGIN to evaluate a range of evaluation metrics. In §4.4.1 we list the untrained metrics we use as well as metrics trained on existing resources, and in §4.4.2 we describe a training set that we designed to train a classifier for the three response categories. We then describe the extent to which these metrics align with the BEGIN categories and analyze the metrics' robustness.

4.4.1 Metrics

Lexical Overlap Metrics This category includes n-gram-based metrics that compare the lexical similarity between the response $\bar{u}_n$ and the knowledge $\mathcal{K}$.[2] We consider BLEU-4[3] [125], ROUGE-L[4] [100], and F1, which measures the word-level lexical overlap between $\bar{u}_n$ and $\mathcal{K}$.

Semantic Similarity Metrics These metrics compare the *semantic* similarity between $\bar{u}_n$ and $\mathcal{K}$. We consider BERTScore [195], which computes the similarity between $\bar{u}_n$ and $\mathcal{K}$ based on the cosine similarity of the sentence embeddings, as well as BARTScore [191] and BLEURT [147]; for implementation details, see Appendix 4.7.

Question-Based Metrics We use Q^2 [70], which computes a factuality score through asking and answering questions. Given a candidate response as input, Q^2

[2]Note that we do not compare the generated responses to the gold responses as they may be unattributable (Sec 4.3.4).

[3]https://github.com/mjpost/sacrebleu

[4]https://github.com/google-research/google-research/tree/master/rouge

generates a corresponding question and identifies potential answer spans in the knowledge source $\mathcal{K}$ that can justify the question–answer pair [32, 172]. It also computes an NLI-inspired similarity score between a candidate response and a predicted answer span in the knowledge source.

Inference-Based Metrics Finally, we study the performance of NLI-based models, trained either on gold NLI benchmarks or on adversarially augmented silver data that we generate. We first describe the metrics trained on gold NLI datasets; we discuss our adversarially augmented dataset (BEGIN-Adversarial) in §4.4.2. We use two transformer-based classifiers: T5-base [133] and RoBERTa-large [102]. We fine-tune them on MNLI [179] and the dialogue inference dataset DNLI [177]. For both datasets, we map the labels (entailment, contradiction, neutral) to the labels (attributable, unattributable, generic) in Begin.

We also train classifiers on AugWow [62], a synthetic dataset designed to evaluate factuality in dialogue systems. This dataset includes three categories: *Supported* responses that are fully verified by $\mathcal{K}$, *Refuted* responses that explicitly contradict $\mathcal{K}$, and responses with *Not Enough Information* (NEI), which do not contain enough information to be verified or refuted by $\mathcal{K}$. We map the labels (supported, refuted, NEI) to the labels (attributable, unattributable, generic) in Begin.

4.4.2 Adversarially Augmented Training Set

This section describes our curated silver training set (BEGIN-Adversarial) for NLI-based attribution classifiers. This dataset includes 8k ($\mathcal{K}$, $\mathcal{H}$, u_p) triples that fit into the three categories: attributable, generic, and unattributable.

Attributable Here we use the original human generated responses u_g from WoW. To avoid human responses that contain opinions or generic chit-chat, we only use response that do not use first-person pronouns and where at least 25% of the words in the response are contained in the evidence.

Unattributable To generate examples that are likely to be unattributable, but are sufficiently challenging to distinguish from attributable ones as to be useful in training a classifier, we use multiple perturbation strategies. First, we directly perturb the knowledge spans $\mathcal{K}$ from the WoW test set and then feed them to GPT2 trained on WoW. We use three perturbation methods, each applied to a different $\mathcal{K}$. First, we swap the subject and the object of $\mathcal{K}$. Second, we replace up to two verbs with verbs of the same tense. Finally, we extract all mentioned entities from different dialogue examples using the SpaCy NER tagger [69], and replace up to two randomly chosen entities in the original $\mathcal{K}$ with entities of the same type. Manual inspection reveals that this usually results in responses that are hallucinations with respect to the original $\mathcal{K}$.

We also generate responses designed to specifically contradict $\mathcal{K}$, using two techniques. First, we directly negate the human response u_g from WoW using the English Resource Grammar parser (ERG; [50]). Second, we replace adjectives in u_g with their WordNet antonyms [112].

Lastly, we gather responses that are off-topic with respect to the information in the $\mathcal{K}$. For a given context, we randomly select a WoW gold response that was based on different $\mathcal{K}$. To avoid easy-to-detect off-topic responses, we sample from conversations that were prompted by the same initial topic word as the target conversation.

Generic Generic responses are generated from the GPT2 model we trained on WoW, using a low softmax temperature of 0.4.

4.4.3 Results

In this section, we report the performance of automatic metrics on the Begin test set.

Lexical and Semantic Metrics The distribution of scores is shown in Figure 4.3. For all metrics, the median score of fully attributable responses is higher than that of generic and unattributable responses, as expected. In many individual cases, however, unattributable responses are scored quite highly, and there is some overlap

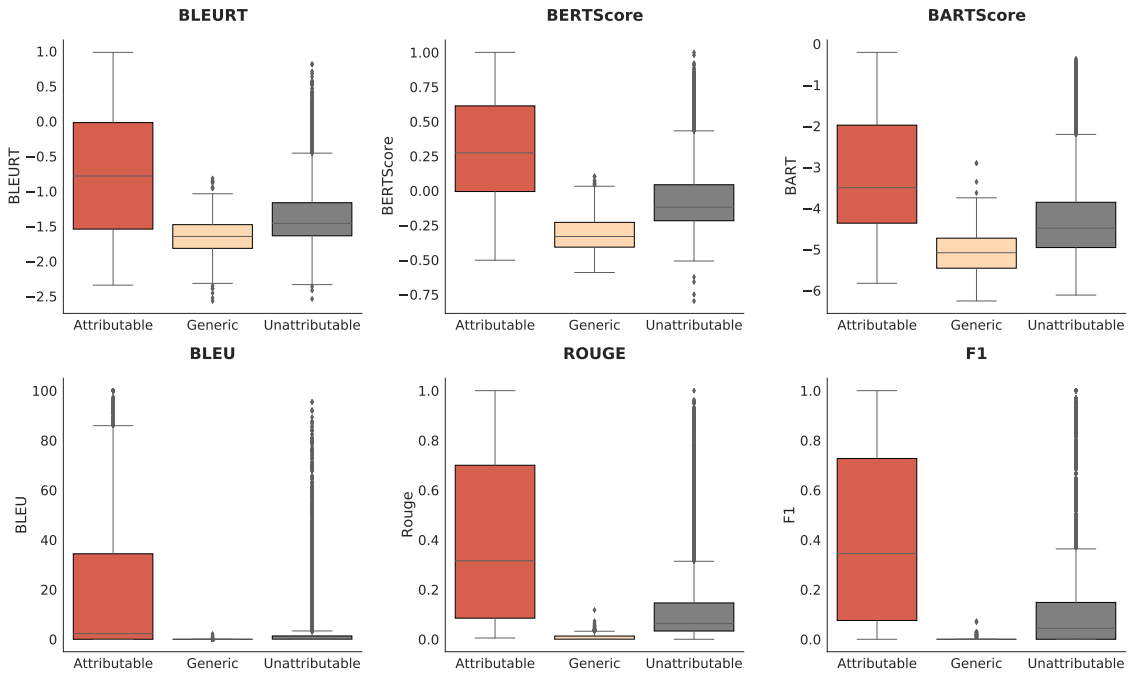


Figure 4.3: The distribution of scores assigned by semantic similarity metrics (upper row) and lexical overlap scores metrics (lower row) to the BEGIN test set.

in the distribution of scores across all three labels, particularly between generic and unattributable responses, indicating that it would be impossible to map these score ranges directly to the BEGIN label taxonomy. Higher scores do not always translate into more desirable response types: Even though a generic response would typically be preferable to an unattributable one in a knowledge-grounded dialogue system, the median scores are lower for generic responses than unattributable ones.

Q^2 Figure 4.4 shows a box plot for each BEGIN class using the Q^2 metric. As in the case of the lexical and semantic metrics, Q^2 scores are typically higher for attributable responses but indistinguishable between generic and unattributable responses.

Inference-Based Classifiers Table 4.2 reports the performance of the NLI-based classifiers on BEGIN. BEGIN-ADVERSARIAL substantially outperforms the classifiers trained on the gold datasets MNLI, DNLI and AugWoW even though it is a significantly smaller resource than those datasets. We also use MNLI as an intermediate fine-tuning

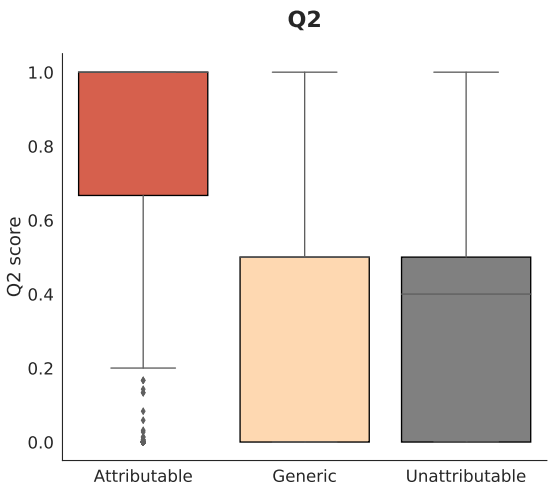


Figure 4.4: The distribution of Q^2 scores for each of the three example categories in the BEGIN test set.

	Test set			Dev set		
Finetuning data	**P**	**R**	**F1**	**P**	**R**	**F1**
T5						
MNLI	48.6	47.9	34.6	52.1	50.7	37.4
DNLI	40.8	56.5	25.6	41.6	59.2	28.6
AugWow	36.8	39.8	37.8	36.7	39.9	38.1
BEGIN-Adv.	46.7	47.4	**45.9**	47.2	47.1	**46.3**
+MNLI	46.9	49.3	45.3	47.6	49.4	46.1
RoBERTa						
MNLI	50.5	51.1	36.4	52.3	53.8	38.5
DNLI	40.2	46.6	27.2	34.9	46.1	29.2
AugWow	41.2	39.2	29.7	29.4	41.4	29.1
BEGIN-Adv.	42.6	46.1	41.1	49.2	45.8	41.1
+MNLI	44.8	45.9	**45.2**	44.9	45.6	**45.1**
Human	96.4	-	-	97.2	-	-

Table 4.2: Precision, recall and F1 of the classifier-based metrics created by fine-tuning T5 and RoBERTa on NLI datasets, AugWow and our adversarial training set. Scores are macro-averaged across labels on the BEGIN test and dev sets.

dataset before fine-tuning on BEGIN-Adversarial.[5] We find that intermediate task fine-tuning can be beneficial when RoBERTa is used as the pretrained model (↑ 4.1 on F1).

Overall, our adversarially generated dataset provides better supervision for detecting our taxonomy than NLI-style datasets. This can be attributed to the fact that NLI-style datasets are designed with a focus on detecting direct contradictions. By contrast, identifying unattributable responses requires detecting multiple types of unverifiable information including, but not limited to, contradictions. At the same time, none of the models exceed 46% F1 score, showing that there is still room for improvement compared to human performance (over 95% precision when comparing human annotations to the majority vote). Finally, T5 and RoBERTa have similar F1 scores despite differences in model size and pretraining corpora, suggesting that simply scaling up the pretrained model may not be sufficient to make progress on this problem.

4.4.4 Are Metrics Measuring Attribution or Extractivity?

Do the metrics perform similarly on both challenging and easier examples? We adopt a density metric from Grusky *et al.* [61] to split the data into three groups—low, medium and high density—based on the extent to which they reuse language from the knowledge sources. Density represents the average length of the text spans in the responses that are copied from the knowledge. Extractive (high density) responses reuse the same phrases as the knowledge source, while abstractive (low density) responses may express the same meaning using a paraphrase.

Results Figures 4.5 and 4.6 show the distributions across different levels of extractivity of the lexical and semantic metrics and the Q^2 score. We observe a common pattern across all metrics: high density responses for all categories (except *generic* on BLEURT) score the highest, followed by medium density and low density responses.

[5]We did not observe a similar improvement when using DNLI as an intermediate task.

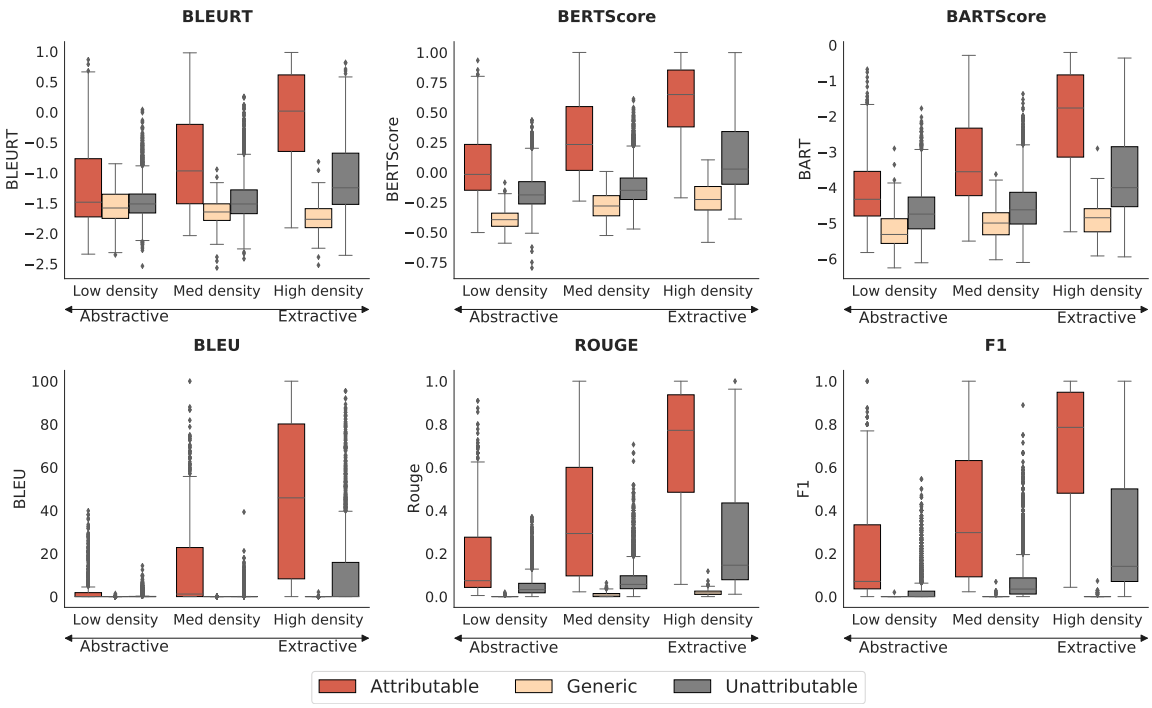


Figure 4.5: Scores assigned to each of the three BEGIN categories by semantic similarity metrics (upper row) and lexical overlap metrics (lower row), broken down by extractivity of the response (the extent to which it copies verbatim from the knowledge).

The differences between the scores of the attributable, generic and unattributable categories are more pronounced in the more extractive responses, and less in the abstractive cases. Only Q^2, though generally unable to separate generic examples, maintains a clear separation between attributable and unattributable examples in the abstractive cases. Moreover, extractivity strongly influences the score assigned to attributable examples; an attributable response is likely to be scored much lower by all of these metrics if it is abstractive. Even more strikingly, unattributable extractive responses score higher on average than attributable abstractive responses in all metrics.

We observe similar trends for the classifiers (Figure 4.7). The performance on classifying attributable responses is much higher in extractive cases than in abstractive ones. In contrast, the performance on unattributable responses is typically worse in the extractive cases. This pattern of results suggests that a response that is unattributable but has a high word overlap with the knowledge is very likely to be

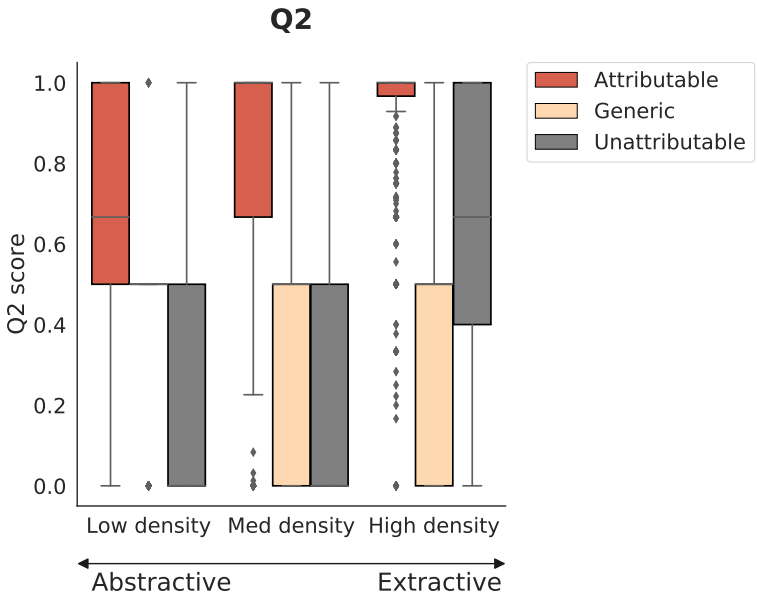


Figure 4.6: Q^2 scores across extractive and abstractive responses on BEGIN test.

misclassified as attributable. In summary, we find that current metrics are relying on the spurious correlation between attribution and word overlap, and do not capture a deep understanding of the notion of attribution (cf. [110]).

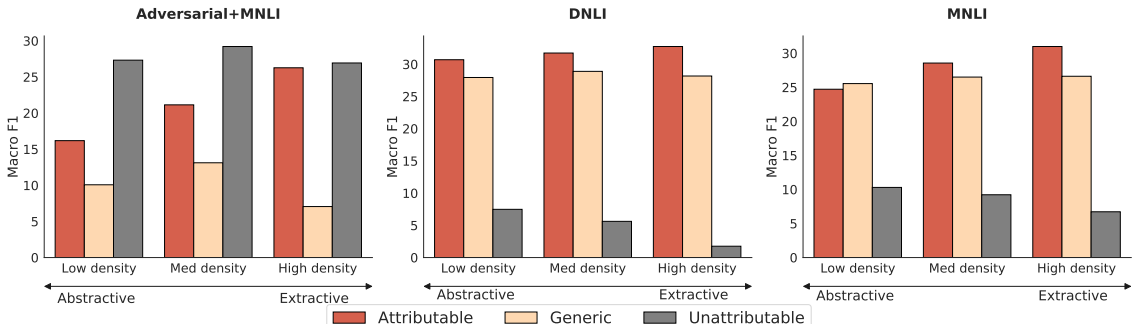


Figure 4.7: Comparison of F1 scores of RoBERTa-based classifiers on BEGIN categories with examples split by density (the extent to which the response copies verbatim from the knowledge).

4.4.5 Robustness to Distribution Shift

We further investigate the robustness of the metrics under distribution shift. Figure 4.8 shows the distributions of both semantic and Q^2 scores across the data broken down

by source. All metrics[6] rate responses from WoW in all categories significantly higher than responses derived from CMU-DoG and TopicalChat. Concerningly, attributable responses generated based on CMU-DoG and TopicalChat receive nearly identical scores to unattributable responses. Likewise, the F1 scores of all the classifiers (Figure 4.9) are higher on the responses from WoW compared to the ones from CMU-DoG and TopicalChat. Specifically, classifiers tested on TopicalChat examples yield the worst F1 scores. For example, RoBERTA-MNLI's F1 score decreases by 10 points when tested on attributable responses from TopicalChat compared to WoW. In general, the metrics appear to perform poorly on datasets that have longer knowledge sources. TopicalChat has on average 271 words in $\mathcal{K}$, followed by CMU-DoG and WoW which have 215 words, 27 words respectively. This shows that shorter knowledge spans correlates with higher metrics performance, pointing to the limited robustness of the metrics.

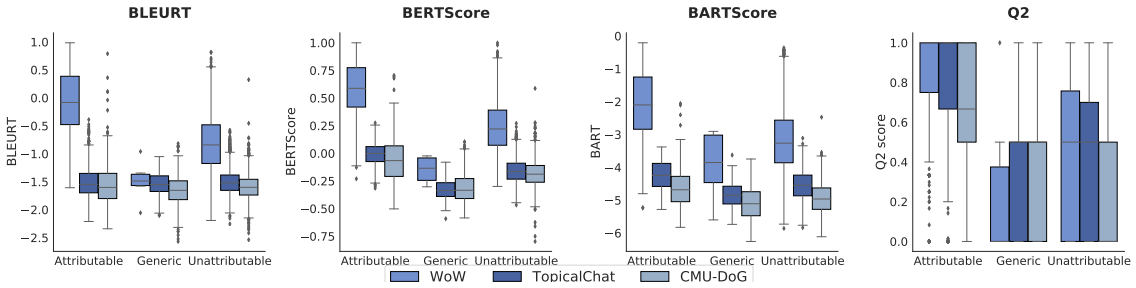


Figure 4.8: Scores of the semantic and Q^2 metrics across the three dialogue corpora we used to train our models.

4.5 Related Work

Analysis of Evaluation Metrics in Natural Language Generation There is extensive interest in analyzing and meta-evaluating neural language generation (NLG) evaluation metrics [53, 54], for various tasks including machine translation [51, 105], data-to-text generation [29], summarization [11, 124, 32, 52, 45, 33], and

[6]We observe similar results for lexical metrics.

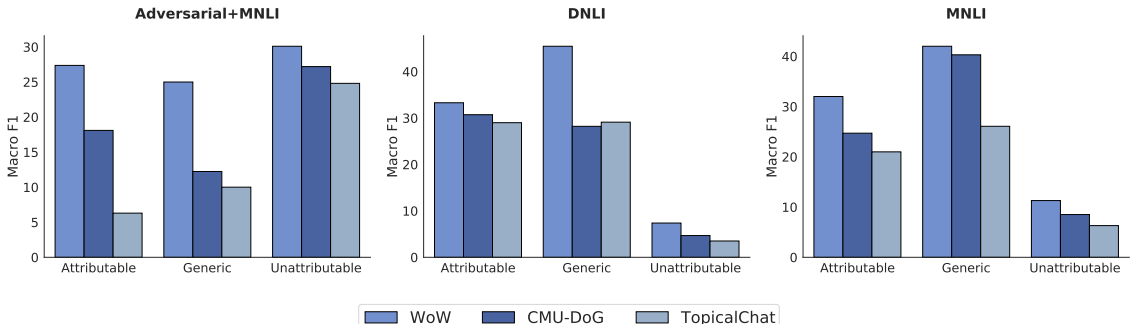


Figure 4.9: Comparison of F1 scores of RoBERTa classifiers on BEGIN categories with examples split by benchmark.

dialogue generation [188, 33]. Most of these studies have compared reference-free and reference-based evaluation metrics to human evaluation. For example, Gabriel *et al.* [52] measured the performance of automated metrics on summaries and compared certain dimensions such as sensitivity and high correlation with human scores. Fabbri *et al.* [45] analyzed metrics in summarization and released human-annotated data for faithfulness across 16 summarization models. We perform a similar meta-evaluation of existing automatic metrics in the context attribution in knowledge-grounded responses. Closest to our work is Durmus *et al.* [33], who found that reference-free evaluation metrics of summarization and dialogue generation rely heavily on spurious correlations such as perplexity and length.

Metrics in Knowledge-Grounded Response Generation In contrast to the significant progress achieved in evaluating many NLG tasks, the evaluation of grounded response generation is a nascent research area [150, 136, 39]. Yeh *et al.* [188] conducted a comprehensive study of existing dialog evaluation metrics. They measured properties such as engagingness and relevance but did not investigate the faithfulness of responses. While hallucination is well-studied in the context of summarization [32, 106, 117, 47], fewer researchers have looked into the problem of assessing hallucination in dialogue systems. Dziri *et al.* [39] introduced a token-level critic that leverages a knowledge graph to identify hallucinated dialogue responses. Rashkin *et al.* [136] proposed a

human evaluation framework to assess output of dialogue models that pertains to the external world and utilized their evaluation framework for conversational QA tasks. Dziri *et al.* [43] introduced a faithful benchmark for information-seeking dialogues and demonstrated that it can serve as training signal for a hallucination critic, which discriminates whether an utterance is faithful or not. An alternative approach for assessing faithfulness uses an auxiliary language understanding task, which measures whether a question answering system produces the same responses for the source document [70]. BEGIN as a testing benchmark should be useful in developing similar metrics further.

NLI and Adversarial Data for Grounded Dialogue Evaluation In this work, we also investigate the performance of classifiers trained on NLI data, extending prior work that has proposed using NLI as a framework for evaluating conversational consistency [176]. Dziri *et al.* [37] also used NLI to evaluate dialogue consistency. They generated a large-scale, noisy synthetic dataset of (premise, hypothesis) pairs tailored for dialogue, based on Zhang *et al.* [193]. We also explore training classifiers on adversarially augmented training data similar to concurrent work from Gupta *et al.* [62] and Kryscinski *et al.* [87], which proposed a synthetic dataset for determining whether a summary or response is consistent with the source document; this dataset was constructed by applying a number of syntactic transformations to reference documents (for a similar approach applied to NLI, see Min *et al.* [113]).

4.6 Implementations

GPT2, T5 We implement these models using the TensorFlow Huggingface Transformers library [184]. During training, we use the Adam optimizer [80] with Dropout [157] on a batch size of 32 with a learning rate of 6.25×10^{-5} that is linearly decayed. The maximum dialogue history length is set to 3 utterances. The model early-stops at epoch {6, 10, 10} respectively for WOW, CMU-DOG and TOPICALCHAT.

CTRL-Dialog We reproduce the results from [135], following the training details in that paper.

DoHA We use the code and the pre-trained model on CMU-DoG that are publicly available by the authors at their Github's account [7]. For WoW and TopicalChat, we follow closely the authors' training procedure described in [129] and we train two models on both datasets.

For each dataset, we save the best model based on the validation set. We use nucleus sampling with $p = 0.9$.

4.7 Model-Based Metrics

Semantic Similarity Models We use BERTScore version 0.3.11. with the DeBERTa-xl-MNLI model [66], which is the recommended model as of the time of investigation. For BLEURT, We use the recommended BLEURT-20 checkpoint [130]. For BARTScore, we use the latest publicly available checkpoint (accessed March 2022) from https://github.com/neulab/BARTScore.

4.8 Conclusion

Contemporary knowledge-based dialogue systems that rely on language models often generate responses that are not attributable to the background knowledge they are expected to convey. In this chapter, we present Begin, a new benchmark to advance research toward robust metrics that can assess this issue. We use Begin to comprehensively evaluate a broad set of existing automatic metrics. We show that these metrics rely substantially on word overlap and fail to properly rank abstractive attributable responses as well as generic responses. They also struggle under distribution shift, assigning low scores to attributable responses grounded on long knowledge sources.

[7] https://bit.ly/3bBup2M

We hope that this work will spur future research on building robust evaluation metrics for grounded dialogue systems.

Chapter 5

FaithDial: A Faithful Benchmark for Information-Seeking Dialogue

A large commonality in the majority of prior work seeks to address hallucination by ameliorating the model [150, 39, 135], but no attempt has been made so far to audit the conversational benchmarks to the best of our knowledge. On one hand, knowledge-grounded conversational benchmarks may contain hallucinations due to error-prone collection protocols, or due to a design framework that encourages informativeness over faithfulness. Existing dialogue systems are typically trained on corpora crowd-sourced through online platforms [30, 57, 114]. With loose incentive to come up with faithfully-grounded utterances on the provided knowledge, crowdworkers may ignore knowledge-snippets altogether, use their personal knowledge or sometimes assume a fictional persona, resulting in conversations that are rife with subjective content and unverified factual knowledge. Figure 5.1 shows a hallucinated conversation from the WoW dataset [30]. On the other hand, neural conversational models are not necessarily designed to generate faithful outputs, but to mimic the distributional properties of the data. This kind of optimization will likely push the models to replicate and even amplify the hallucination behaviour at test time [8]. The presence of even few hallucinated responses may skew the data distribution in a way that curbs the model's ability to generate faithful responses [77]. In this chapter, we investigate the root causes of hallucinations and we introduce a new hallucination-free dialogue

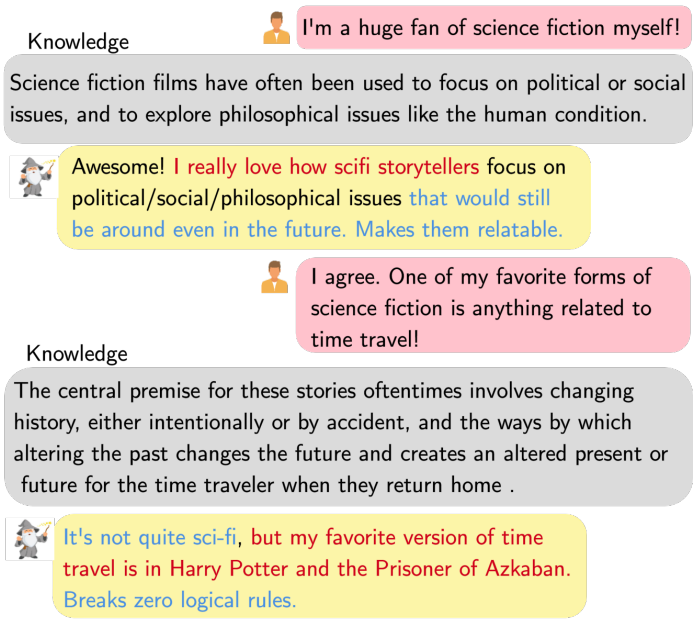


Figure 5.1: An example of a hallucinated conversation from the Wizard of Wikipedia dataset [30]. The wizard (yellow) is hallucinating information that cannot be inferred from the knowledge-snippet: hallucinated subjective content (red) and hallucinated objective content (blue).

benchmark.

5.1 On the Origin of Hallucinations in Conversational Models

5.1.1 Hallucinations in Benchmarks

We conduct a human study on three English crowdsourced knowledge-grounded conversational benchmarks: Wizard of Wikipedia (WoW), CMU-DoG and TopicalChat. These datasets consist of dialogues between two speakers, where the goal is to communicate information about particular topics while speakers are presented with a knowledge snippet relevant to the current turn.

Wizard of Wikipedia (WoW) Details about the dataset are discussed in Section 2.4.6. We omitted examples where the Wizard did not explicitly select a passage as evidence for the response or where there was no dialogue history. We also use the

"unseen" topic portion of the test data. Overall, we used 82722 training examples, 8800 development examples, and 3902 test examples.

CMU-DoG The CMU-DoG dataset [201] consists of conversations about films. Each response is expected to be grounded in a section from Wikipedia. Workers can have either asymmetric or symmetric roles. In the asymmetric setting, one worker is asked to persuade the interlocutor to watch the movie using arguments from the document where only the persuader has access to the document. In the symmetric role, workers discuss together the content of the document. In total, there are 78136, 13800 and 13796 grounded responses (training/dev/test).

TopicalChat TopicalChat [57] consists of dialogues about a variety of topics. Workers are provided relevant facts from Reddit, Wikipedia and news articles. Analogous to CMU-DoG, the data collection protocol consists of two scenarios. In the symmetric scenario, workers have access to the same knowledge source; in the asymmetric scenario, they have access to different sources. They are asked to use the information from the documents to chat knowledgeably about the topic. In total, the dataset has 134572, 8790 and 8081 grounded responses (training/dev/test).

Response Classification Taxonomy Following the definitions of the BEGIN taxonomy [42] and the AIS framework [136] of evaluating response attribution, we annotate each response based on whether it can be inferred exclusively from the knowledge-snippet as follows: **Entailment**: a response is fully supported by the knowledge, i.e., any information it contains must be attributed to the knowledge. **Hallucination**: a response's factual correctness cannot be fully verified from the knowledge-snippet (even if it is true in the real world). More specifically, personal opinions, experiences, feelings, internal assessments of reality that cannot be attributed to the information present in the source document, are considered hallucinations. **Partial Hallucination**: part of the response is hallucinated while the rest is entailed by the

VRM Type	Description	Example
Disclosure	Reveal the speaker's subjective opinions, personal experience, thoughts, feelings, wishes, and intentions.	*"I think science fiction is an amazing genre. Future science, technology they're all interesting."*
Edification	Concerns information that is, in principle, objective.	*"Recycling includes items like metal and plastic."*
Advisement	Corresponds to guiding the behaviour of the addressee through: commands, requests, suggestions, advice, permission, prohibition.	*"You should be patient and persistent to succeed."*
Confirmation	Compares the speaker's experience with the other's by expressing shared ideas/memories/beliefs, or by agreement/disagreement	*"I agree that love encompasses a variety of different emotional and mental states."*
Question	Concerns requesting information or guidance.	*"What is your favorite song?"*
Acknowledge	Expresses no content, it conveys only receipt of communication from the other's speaker.	*"Mmm. OK,..."*, *"Yeah, ..."*, *"Hello, ..."*

Table 5.1: The definitions of the VRM types with examples.

source knowledge. **Generic**: a response that is vague and does not convey any factual information such as "*Sounds good*" or "*I'm not sure about that*". **Uncooperative**: an entailed response that does not follow the principles of conversational cooperation according to Gricean maxims [59]. The response may be purposefully misleading, or showing a general unwillingness to cooperate with the interlocutor, resulting in an incoherent communication.

To understand the linguistic nature of hallucinations, we further annotate responses based on a linguistic coding system for discourse phenomena, dubbed Verbal Response Modes (VRM; [158]). Concretely, we label a turn with the following speech acts: **Disclosure**, **Edification**, **Advisement**, **Confirmation**, **Question** and **Acknowledgement (Ack.)**. Table 5.1 displays the definition for each VRM type with examples. We opted for the VRM taxonomy as it offers a simple way of codifying responses into categories that are sufficient for our analysis whereas one can also opt for a more demanding annotation scheme [15].

	BEGIN	VRM
CMU-DoG	0.85	0.78
TopicalChat	0.83	0.72

Table 5.2: Fleiss Kappa Scores on 200 train Human-Human responses from the CMU-DoG and TopicalChat benchmarks.

5.1.2 Human Evaluation Study

We follow a two-stage annotation protocol where we first ask two experts to judge the attribution of 200 randomly sampled train responses with respect to the source knowledge. The experts were students with linguistics background, fluent in English, and were trained for the task by exchanging rigorous discussions with the authors. As part of this stage, they were required to write justifications for 50 samples articulating the reasoning for the provided ratings. The collected justifications were helpful in understanding the reasoning used to reach their ratings and in laying the groundwork for designing the second round of annotations. For inter-annotator agreement, we measure Fleiss' Kappa scores on both BEGIN and VRM. WoW achieved 0.89 on BEGIN and 0.78 on VRM, indicating substantial agreement. Annotations on CMU-DoG and TopicalChat achieved nearly similar agreement (See Table 5.2). The high agreement scores align with the findings in AIS on WoW [136].

The second round corresponds to a large-scale annotation of 4K randomly sampled train responses using non-expert annotators from AMT. This round is crucial to ensure that the obtained results from the experts are reliable enough to draw conclusions about the quality of the data. As human annotation is expensive, we perform the non-expert annotations only on the WoW benchmark while restricting ourselves to expert annotations on CMU-DoG and TopicalChat data. We choose WoW over the other two datasets as the source knowledge is more amenable to faster annotation (TopicalChat: 300 words > CMU-DoG: 215 words > WoW: 27 words). Below, we detail our AMT task design and how we ensure data quality:

Task Design To streamline the process for raters we break down the task into hierarchical (yes/no) questions. We summarize this procedure below, and provide the exact questions in the following paragraphs. First, we ask annotators to judge whether the response contain information that is not supported by the source. If yes, we ask them to indicate the type of the unsupported information (e.g., unsupported opinion, unsupported fact, etc). In a followup question, we ask them to indicate whether there are any supported information besides the hallucinated content. If the response was not hallucinated, we present them with two follow-up questions about whether the response is entailing the source or generic. Finally, if the response entails the source, we ask whether it is coherent with the history.

AMT Data Quality To access the initial staging round in AMT, workers have to pass a qualification test by answering correctly 14 questions about BEGIN and VRM. Moreover, they had to be situated in the United States and Canada. Before being granted access to the main annotation task, workers would have access only to a small pilot round (batch size $\sim$ 50 HITs). In this round, we carefully inspect each of the workers annotations for adherence to the instructions, and provide feedback via email to those who committed errors.

At the end of this round, we revoke access for workers who provide poor quality annotations. Next, we launch the main annotation stage which is larger (batch size $\sim$ 400 HITs). We perform daily manual inspection and we send detailed feedback to workers who commit persistent error patterns. We reject poor quality work in this stage and repeated rejections lead to blocking the workers from the task indefinitely. In total, we ended up with 4 workers annotating the 4k responses. The workers were informed that their annotations would be used for research purposes and their workers ID would be anonymous when we release the data.

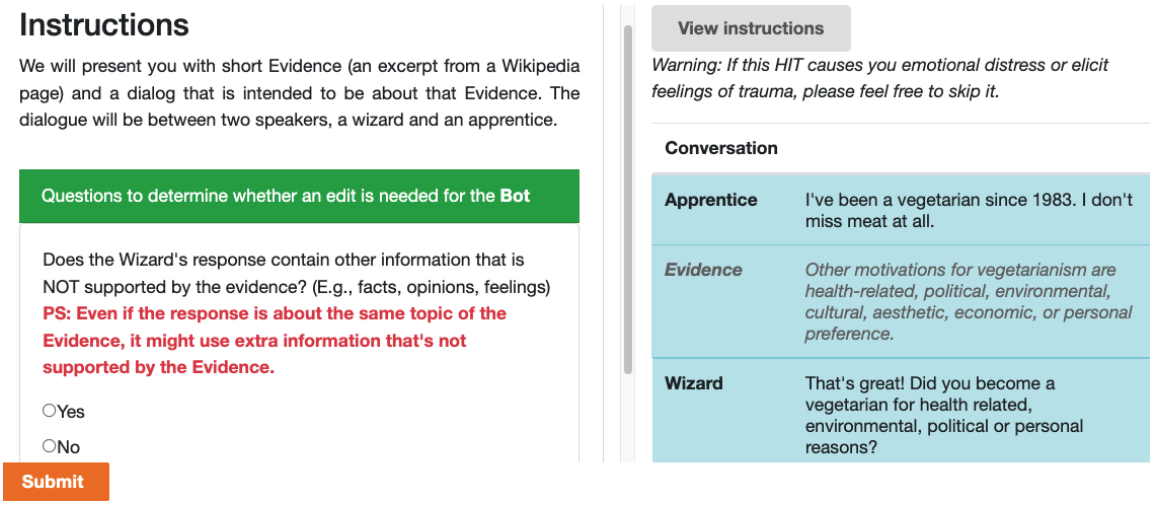


Figure 5.2: AMT Annotation interface for determining BEGIN and VRM classes (1)

AMT Human Instructions AMT Human annotation interfaces are depicted in Figure 5.2 and Figure 5.3. We pay workers an hourly wage around 18-20 USD which is above the minimum wage rate. Workers were asked the following questions:

1. Does the Wizard's response contain other information that is NOT supported by the evidence? (E.g., facts, opinions, feelings)?

   (a) If the response is hallucinated, what is the type of the unsupported information? (expressing a personal experience, expressing an opinion, expressing feelings, expressing unsupported facts, giving advice, acknowledging with information from the human)

   (b) Besides unsupported information, does the Wizard's response contain thoughts/opinions/feelings/facts that are supported by the Evidence?

2. If the response is not hallucinated, is it faithful to the source or generic? (Faithful, Generic)

3. If the response if faithful, is it cooperative with the Human's response?

In total, we selected 4 trusted workers to annotate the 4k responses. To compute the inter-annotator agreement, we assign three workers per response in a secondary

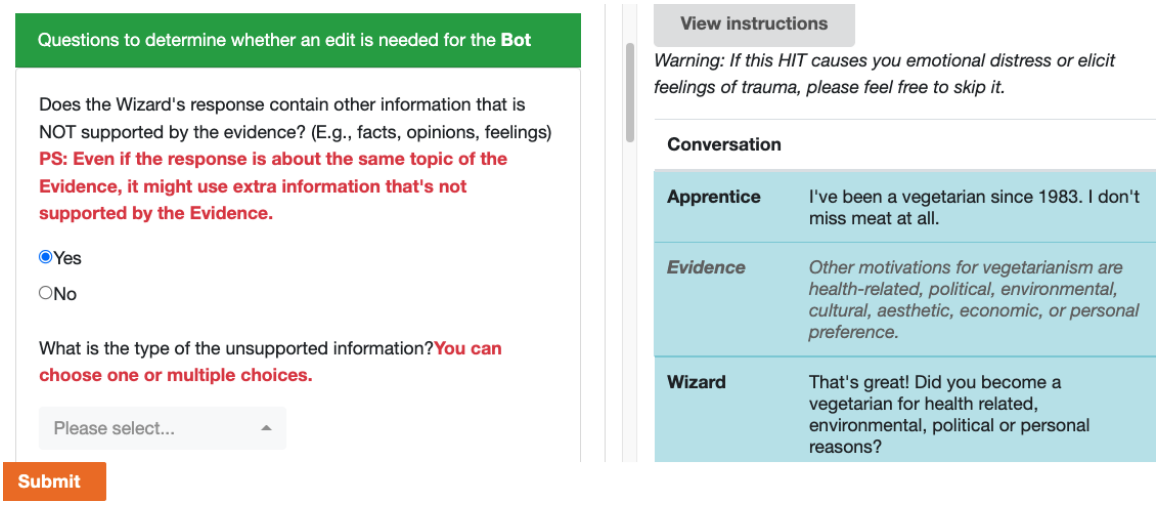


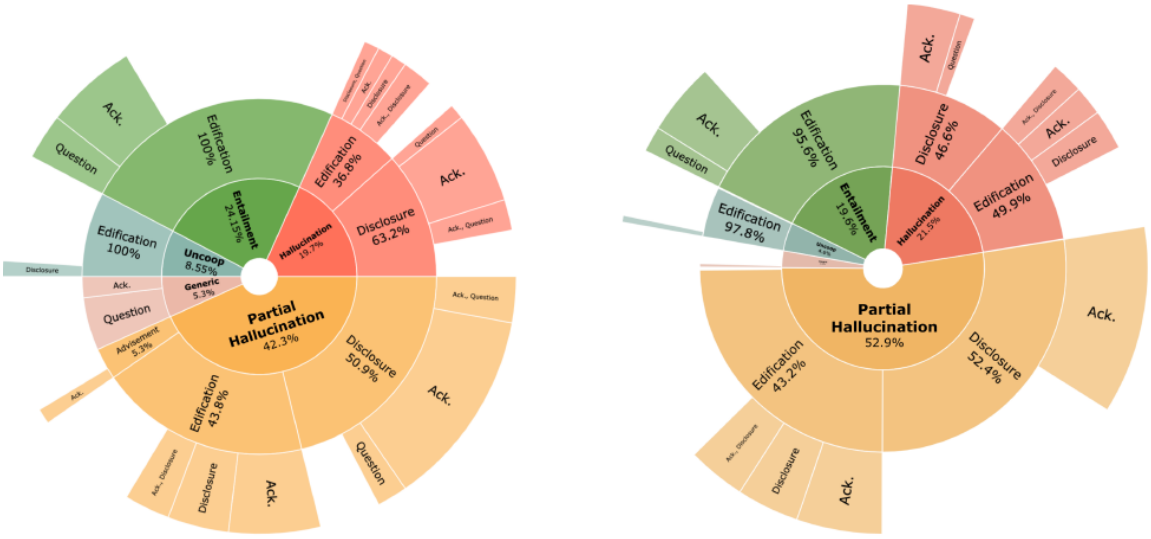
Figure 5.3: AMT Annotation interface for determining BEGIN and VRM classes (2)

task, and ask each of them to judge 500 responses. Reported Fleiss' Kappa agreements were 0.75 for BEGIN and 0.61 for VRM. Although substantial, the agreement is lower than the experts' one and this is expected as they have stronger linguistic background.

5.1.3 Human Study Results

We seek to answer the following questions:

(Q1) How much hallucination exists in the benchmarks? Figure 5.4 shows the breakdown of each BEGIN categoty in WoW and compares expert annotations versus AMT workers. Surprisingly, WoW is fraught with hallucinations. Expert annotations on 200 responses show that hallucinated responses are largely mixed with faithful content (42.3% v.s. 19.7% fully hallucinated responses), which amounts to 62% hallucinations in total. These results generalize even on larger data; we can see that the portion of hallucinated responses increased to 74.4% when evaluated on 4K samples. Our analysis shows similar trends on the CMU-DoG and TopicalChat benchmarks (Figure 5.5). CMU-DoG contains 61.4% responses that are purely hallucinated against only 16.2% responses that are fully entailing the source knowledge and TopicalChat has similar results (63.9% hallucination v.s. 22.9% entailment).



(a) Expert annotations (200 responses)

(b) Non-expert annotations (4000 responses)

Figure 5.4: A representative FAITHDIAL annotation: subjective and hallucinated (red) information present in the wizard's utterance of WoW data are edited into utterances faithful to the given knowledge (green). In FAITHDIAL, the wizard assumes the persona of a bot.

Exemplars of hallucinated responses are depicted in Table 5.15. These findings raise the question on the quality of dialogue datasets.

(Q2) What are the hallucination strategies used in human-human data? Figure 5.4 and Figure 5.5 show the VRM breakdown for each BEGIN category in the three benchmarks. We make the following observations: The majority of hallucinations belong to *disclosure* (i.e., subjective information) in all benchmarks (50.9%, 56.2% and 61.5% in WOW, CMU-DOG and TOPICALCHAT respectively). Although the strategy of sharing subjective information such as thoughts, opinions and feelings is natural in conversations, it often comes at a cost of ignoring the knowledge snippet in these datasets. Moreover, *edification* is also a common phenomenon in hallucinated responses, suggesting that humans not only discuss subjective information but also bring extra unsupported facts, either true or false. Other linguistic modes are also associated with hallucinations such as acknowledging unsupported claims or asking irrelevant questions. Conversely, entailment responses have high percentage of

(a) CMU-DoG responses

(b) TopicalChat responses

Figure 5.5: BEGIN and VRM breakdown of gold responses from CMU-DoG and TopicalChat. The inner circle shows the breakdown of BEGIN classes and the outer shows the VRM types in each BEGIN type: Hallucination (red), Entailment (green), Partial Hallucination (yellow), Generic (pink), and Uncooperative (blue).

edification ($> 70\%$) with information inferred from the knowledge snippet.

5.1.4 Hallucination Amplification in Models

Next, we investigate how much models amplify the hallucination phenomenon at inference time. We consider a range of representative models:

• **GPT2** [132, 183] is an autoregressive model which takes as input a concatenation of the knowledge and the history.

• **DoHA** [129] builds a BART-based conversational model [93] for knowledge-grounding, with a two-view attention mechanism to handle separately the encoded document and the history during generation.

• **CTRL** [135] augments the GPT2 model with control tokens [78] that guide the generation towards less subjective and more entailed content.

We fine-tune each model on the benchmarks and use nucleus sampling [68] with $p = 0.6$ for decoding.

Data	Knowledge & History	Responses
WoW	**Knowledge**: Cinematography (also called "Direction of Photography") is the science or art of motion-picture photography by recording light or other electromagnetic radiation, either electronically by means of an image sensor, or chemically by means of a light-sensitive material such as film stock. **History**: Hi buddy, What you think about cinematography?	**GPT2**: Cinematography is another title I'm currently in school for. I've been doing the art of photography for over 15 years now. (Hallucination) **DoHA**: I do enjoy the science and art of motion-picture photography. (Partial Hallucination) **CTRL**: You're right. Cinematography is an art of motion -picture photography by recording light or other electromagnetic radiation, either electronically by means of an image sensor, or chemically by means of a light-sensitive material such as film stock. What about you ? (Entailment)
CMU-DoG	**Knowledge**: Mean Girls is a 2004 American teen comedy film directed by Mark Waters and written by Tina Fey. The film is partially based on Rosalind Wiseman's 2002 non-fiction self-help book Queen Bees and Wannabes, which describes female high school social cliques and the damaging effects they can have on girls. Genre: Comedy. MovieName: Mean Girls rating: Rotten Tomatoes: 84 % and average: 6.9/10. **History**: I think Rachel McAdams had an even better role as Regina George however!	**GPT2**: Mean Girls was a classic. (Hallucination) **DoHA**: Yes I agree! she was great and she was better than Mark Waters! (Hallucination) **CTRL**: I see. Mean Girls is a comedy movie that was directed by Mark Waters in 2004. (Entailment)

Table 5.3: Generated responses from different models based on Wizard of Wikipedia [30] and CMU-DoG [201] test samples.

Implementation Details

GPT2: This model was implemented using the Pytorch Huggingface Transformers library [184] and the Pytorch-lightning library[1]. To train the models, we use the Adam optimizer [81] with Dropout [157] on a batch size of 32 with a learning rate of 6.25×10^{-5} that is linearly decayed. The maximum dialogue history length is set

[1]https://github.com/PyTorchLightning/pytorch-lightning

	Model	R-L↑	Hallucination Rate↓			Entailment Rate↑		
			Full	Partial	Overall	Entail.	Uncoop.	Overall
WoW	Gold	36.1	19.7	42.3	62.0	24.1	8.5	32.7
	GPT2	27.0	66.0	15.2	81.2	11.7	3.6	15.3
	DoHA	30.6	39.6	28.9	68.5	12.7	7.1	19.8
	CTRL	51.3	31.0	5.0	36.0	19.5	42.0	61.5
CMU-DoG	Gold	4.1	61.4	5.1	66.5	16.2	4.1	20.3
	GPT2	4.6	75.5	6.0	81.5	5.5	5.5	11.0
	DoHA	5.1	62.5	10.0	72.5	8.5	5.0	13.5
	CTRL	6.9	62.5	4.5	67.0	13.5	17.0	30.5
Topical	Gold	1.2	46.8	17.1	63.9	22.9	0.5	23.4
	GPT2	6.9	70.5	8.5	79.0	6.5	5.0	11.5
	DoHA	4.0	53.0	25.0	78.0	9.0	5.0	14.0
	CTRL	7.9	48.5	16.7	65.2	12.1	20.7	32.8

Table 5.4: Amplification of models on the test data from WoW and CMU-DoG and TopicalChat. 'Entail.' and 'Uncoop.' mean entailment and uncooperative, respectively. R-L measures the ROUGE-L scores between the response and the knowledge.

to 3 utterances. The model early-stops at epoch {7, 8, 8} respectively for WoW, CMU-DoG and TopicalChat. The average runtime is {1.5, 3, 3} hours for WoW, CMU-DoG and TopicalChat respectively.

DoHA: We use the pre-trained model on CMU-DoG that is publicly available[2]. However, since no models trained on WoW and TopicalChat have been released, we follow closely the training procedure described in Prabhumoye *et al.* [129] and we train two models. The average runtime of these models is {5, 10} hours for WoW and TopicalChat respectively.

CTRL: We implement the model ourselves since the code and the model were not released by the authors. We follow training details in Rashkin *et al.* [135] and implement this model using the Pytorch Huggingface Transformers library and the Pytorch-lightning library. Additionally, we had multiple discussions with the authors to make sure that our implementation is accurate.

[2]https://bit.ly/3bBup2M

We save the best model based on the validation set, for all datasets. Training for all models is done on an Nvidia V100 GPU 32GB and for inference, we use nucleus sampling with p=0.6.

As seen in Table 5.4, CTRL is the best model followed by DoHA based on the hallucination ratio. Table 5.3 shows a sample of generated responses. Similar to the analysis in §5.1.1, we task the same two linguists to analyze model-generated responses for 200 randomly-selected test samples from each benchmark.

(Q3) Do state-of-the-art conversational models amplify hallucination? Table 5.4 shows the degree of amplification across different models trained on the three benchmarks. Numbers report the percentage of each class in the data. Contrasting this with human gold responses, the models not only hallucinate but also amplify the percentage of hallucinations, except CTRL on WoW. For example, GPT2 amplifies full hallucination by 19.2% in WoW, 15% in CMU-DoG and 15.1% in TopicalChat. Conversely, it reduces entailment by 17.4%, 9.3% and 11.9% respectively. This suggests that hallucination patterns are easier to learn than entailment. Among the three, CTRL hallucinates the least at the expense of producing a high number of uncooperative responses. Although these responses are entailing the knowledge, they are not coherent with the history. A closer inspection shows that most uncooperative responses are extractive, i.e., they copy big chunks of the evidence without adapting the content to the history or they just output an exact copy of the entire evidence. This is also reflected in high ROUGE scores between the response and the knowledge, corroborating the extractive nature of CTRL compared to the gold responses. This behavior is not surprising as CTRL was optimized to maximize the overlap with the knowledge. Overall, these results demonstrate that hallucination is not only a reflection of training data issues, but also a consequence of the weaknesses of models.

We hypothesize that there are multiple factors that can contribute to the models' deficiencies: First, the exposure bias [134] caused by teacher forcing can make hal-

lucination worse as the model may over-rely on previously predicted words which in turn can aggravate error propagation. Second, maximum likelihood estimation can be fragile to noisy data points as it necessitates models to assign high probability mass to all test references, resulting in unstable behavior—a fact observed in machine summarization [77]. Moreover, we link this issue to the decoding strategies used at test time. We conjecture that models—when conditioned on factual knowledge—often assign the highest probability mass to the correct response and sampling based on other distributions (e.g. top-k or nucleus) may invite hallucination in the generation process. And lastly, we hypothesise that the behavior of these models is ultimately shaped by the bias learned from internet text during pre-training [115]. We leave investigating the role of each factors to hallucination amplification for future work.

(Q4) What are the hallucination strategies used by models? Surprisingly, different models use different strategies for hallucination. While DoHA and GPT2 predominantly rely on and amplify *disclosure*, CTRL relies on *edification*. This is because CTRL is trained explicitly to avoid pronouns (a crucial ingredient for disclosure) and to generate entailed responses. As a side-effect, it ends up amplifying uncooperative responses (by 33.5%, 12.9% and 20.2% in WoW and CMU-DoG as seen in Table 5.4).

5.2 FaithDial: Introduction

In the previous section, we investigated the underlying roots of hallucination and found that the gold-standard conversational datasets [30, 57, 201]—upon which the models are commonly fine-tuned—are rife with hallucinations, in more than 60% of the turns. An example of hallucination in Wizard of Wikipedia (WoW; [30]) is shown in the red box of Figure 5.6. In WoW, an information SEEKER aims to learn about a topic and a human WIZARD harnesses knowledge (typically a sentence) from Wikipedia to answer. This behavior, where the human WIZARD ignores the knowledge

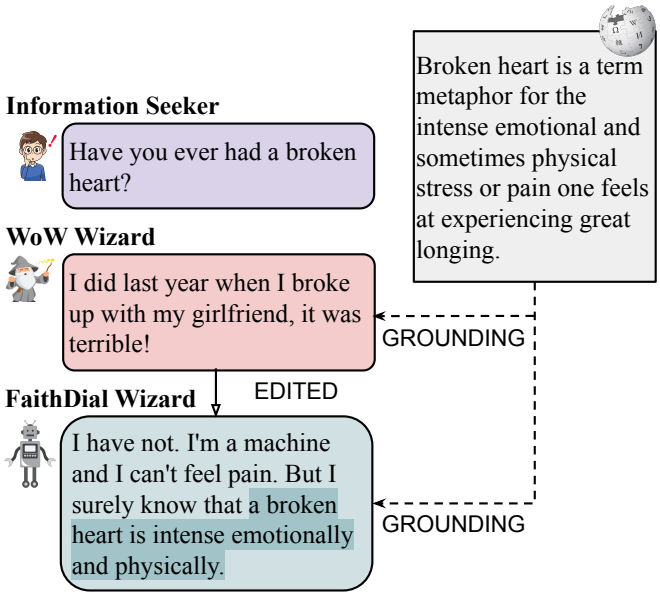


Figure 5.6: A representative FAITHDIAL annotation: subjective and hallucinated (red) information present in the wizard's utterance of WoW data are edited into utterances faithful to the given knowledge (green). In FAITHDIAL, the wizard assumes the persona of a bot.

snippet and assumes a fictitious persona, can later reverberate in the dialogue system trained on this kind of data. Instead, the ideal WIZARD response, highlighted in green, should acknowledge the bot's nature, and whenever the knowledge is not sufficient or relevant, it should acknowledge its ignorance of the topic.

Unfortunately, modeling solutions alone cannot remedy the hallucination problem. By mimicking the distributional properties of the data, models are bound to 'parrot' the hallucinated signals at test time [8]. What is more, we observe that GPT2 not only replicates, but even amplifies hallucination around 20% when trained on WOW. This finding also extends to models that are designed explicitly to be knowledge-grounded [129, 135]. Filtering noisy or high-error data [77] is also prone to failure as it may either break the cohesion of discourse or it may require excluding entire dialogues.

In this section, we adopt instead a data-centric solution to address hallucinations and create FAITHDIAL, a new benchmark for faithful[3]knowledge-grounded dialogue.

[3]Faithfulness is sometimes referred to as attribution [42, 136] and fidelity [152].

Specifically, we ask annotators to amend hallucinated utterances in WoW by making them faithful to the corresponding knowledge snippets from Wikipedia and acknowledging ignorance when necessary. This approach is vastly more scalable than creating FaithDial from scratch while retaining the cohesiveness of conversations. Moreover, it allows us to shed light on hallucinations by contrasting corresponding wizard's responses in WoW and FaithDial. As a result, FaithDial contains around 50K turns across 5.5K conversations. Extensive human validation reveals that 94.4% of the utterances in FaithDial are faithful (i.e., without hallucinations), compared to only 20.9% in WoW. Moreover, we benchmark several state-of-the-art models [132, 143, 133, 135] on dialogue generation. If trained on FaithDial, we find that they are significantly more faithful while also enhancing other dialogue aspects like cooperativeness, creativity, and engagement. These benefits also generalize to other knowledge-grounded datasets like CMU-DoG [201] and TopicalChat [57] in a zero-shot transfer setting.

FaithDial also provides supervision for hallucination critics, which discriminate whether an utterance is faithful or not. We source positive examples from FaithDial and negative examples from WoW. Compared to other dialogue inference datasets [178, 120], the classifiers trained on this data (which we call FaithCritic) transfer better to general NLU tasks like MNLI [179] and achieve state-of-the-art on dialogue-specific knowledge grounding benchmark (BEGIN) [42] in a zero-shot setting.

Thus, FaithDial holds promise to encourage faithfulness in information-seeking dialogue and make virtual assistants both more trustworthy. We will release data and code for future research.

5.3 FaithDial: Dataset Design

Given the motivations adduced above, the primary goal of this work is to create a resource for faithful knowledge-grounded dialogue that allows for both training high-quality models and measuring the degree of hallucination of their responses.

Definition 5.3.1 (Faithfulness)**.** *Given an utterance u_n, a dialogue history $\mathcal{H} = (u_1, \ldots, u_{n-1})$, and knowledge $\mathcal{K} = (k_1, \ldots, k_j)$ at turn n, we say that u_n is faithful with respect to $\mathcal{K}$ iff the following condition holds:*

- *$\exists\, \Gamma_n$ such that $\Gamma_n \vDash u_n$, where $\vDash$ denotes semantic consequence and Γ_n is a non-empty subset of $\mathcal{K}_n$. In other words, there is no interpretation $\mathcal{I}$ such that all members of Γ_n are true and u_n is false.*

Hence, an utterance can optionally be grounded on multiple facts but not none.

In what follows, we present the design of our task as well as our annotation pipeline to curate FaithDial. In our dialogue setting, we simulate interactions between two speakers: an information seeker and a bot wizard.

Definition 5.3.2 (Information Seeker: A Human)**.** *The information* seeker*, a human, aims at learning about a specific topic in a conversational manner. They can express subjective information, bring up a new set of facts independent from the source $\mathcal{K}$, and even open up new sub-topics.*

From the perspective of Definition 5.3.2, utterances pronounced by the seeker have a large degree of freedom. For example, the human can chat about personal life and can ask a diverse set of questions. On the other hand, the wizard is more restricted on what they can communicate.

Definition 5.3.3 (Wizard: A Bot)**.** *The Wizard, a bot, aims at conversing in a knowledgeable manner about the* seeker*'s unique interests, resorting exclusively to the available knowledge $\mathcal{K}$. They can reply to a direct question or provide information about the general topic of the conversation.*[4]

From Definition 5.3.3, it follows that there are three key rules the bot must abide by: first, it should be truthful by providing information that is attributable to the

[4]To encourage naturalness in the response, annotators were also asked to express empathy such as "*I'm sorry about ...*". in case the Seeker expresses a very unfortunate event.

Dataset	Generic	Hallucination		Entailment	
		Full	Partial	Faith.	Uncoop.
WoW	5.3	19.7	42.3	24.1	8.5
CMU	13.2	61.4	5.1	16.2	4.1
Topical	12.7	46.8	17.1	22.9	0.5

Table 5.5: The breakdown of responses from WoW, CMU-DoG and TopicalChat according to BEGIN taxonomy [42]. "Faith." refers to faithful responses and "Uncoop." refers to faithful but uncooperative responses given the conversation history.

source $\mathcal{K}$. Second, it should provide information conversationally, i.e., use naturalistic phrasing of $\mathcal{K}$, support follow-up discussion with questions, and prompt user's opinions. Third, it should acknowledge its ignorance of the answer in those cases where $\mathcal{K}$ does not include it while still moving the conversation forward using $\mathcal{K}$.

5.3.1 Data Selection

Rather than creating a novel benchmark from scratch, however, we opt for fixing problematic utterances (which are the majority) in existing dialogue benchmarks [40]. The reason is three-fold: 1) while mostly hallucinated, existing datasets still contain useful faithful information. 2) as correction is faster than creation from scratch, this enables us to annotate examples on a larger scale; 3) two versions of the same dialogue turn, either hallucinated or faithful, can provide signal for (contrastive) learning and evidence for a linguistic analysis. In particular, we focus on WoW as our benchmark backbone.

Initial pilot study revealed that WoW dialogues are more suitable for editing compared to other prominent knowledge-grounded dialogue benchmarks: TopicalChat [57] and CMU-DoG [201]. In fact, as shown in 5.5, WoW is relatively less hallucinated compared to CMU-DoG and TopicalChat. Moreover, full hallucinations—responses that contain no faithful content and that therefore need to be entirely thrown out—are highly prevalent in the latter two (61.4% in CMU-DoG and 46.8% in TopicalChat and only 19.7% in WoW). Moreover, knowledge snippets in WoW tend to be shorter, which is preferable as longer knowledge is correlated with increased hallucination

due to the constrained cognitive capacity for text navigation and comprehension in humans [25, 26].

Our first step consists in filtering out WoW conversations where ground-truth knowledge $\mathcal{K}$ was not given, and annotators relied on personal knowledge instead. Then, we focus on SEEKER-initiated conversations and sample 44% from the train (4094 conversations) and 100% from validation (764 conversations) and 100% from test (791 conversations)[5].

5.3.2 Crowd-sourced Annotations

WoW	**FaithDial**
Seeker: I would love to learn to surf but I'm terrified of sharks! Do you surf?	**Seeker**: I would love to learn to surf but I'm terrified of sharks! Do you surf?
Knowledge: Surfing is a surface water sport in which the wave rider, referred to as a surfer, rides on the forward or deep face of a moving wave, which is usually carrying the surfer towards the shore.	
Wizard: I absolutely love to surf, just riding on the forward face of a moving wave is so exhilarating. Are there a lot of sharks near you? (BEGIN: Partial hallucination, VRM: disclosure)	**Wizard**: Unfortunately, it's impossible for me to surf. Remember, I'm just your virtual bot! Sorry to hear you're terrified from sharks. However, did you know that a surfer is someone who can ride on either the forward or deep face of a wave which typically sends her/him close to the shore?
Seeker: I have no idea how many sharks are in the Gulf of Mexico but there out there! lol Where do like to surf? (incoherent with the freshly edited Wizard response)	**Seeker**: Yeah, what scares me is the sharks, is there any non-ocean waves out there?
Knowledge: Waves suitable for surfing are primarily found in the ocean, but can also be found in lakes or in rivers in the form of a standing wave or tidal bore.	
Wizard: There's a little hidden cove near me that has great waves, although waves suitable for surfing are primarily found in the Pacific ocean. Have you ever watched people surf? (BEGIN: hallucination, VRM: disclosure & edification)	**Wizard**: For sure there is. An alternative to ocean waves could be lakes and rivers which have a phenomenon called standing waves. Have you ever watched people surf?

Table 5.6: A dialogue example showing the process of editing WoW utterances to convert them to FaithDial utterances. Text highlighted in red indicates hallucinated content. Text in violet indicates the BEGIN labels and the speech act VRM labels as identified by annotators.

Following the guidelines for ethical crowdsourcing outlined in Sheehan [149], we hire

[5]We use the original WoW splits.

Edit Type	Percentage
Wizard edits	84.7%
Seeker edits	28.1%
Wizard edits per conversation	3.8 turns
Seeker edits per conversation	1.2 turns

Table 5.7: Amendment statistics of WoW

Amazon Mechanical Turk (AMT) workers to edit utterances in WoW dialogues that were found to exhibit unfaithful responses. First, workers were shown dialogues from WoW and asked to determine whether the WIZARD utterances are faithful to the source knowledge. To guide them in this decision, they were additionally requested to identify the speech acts (VRM taxonomy; Stiles [158]) such as disclosure, edification, question, acknowledgment, etc; and the response attribution classes (BEGIN taxonomy; Dziri *et al.* [42]) such as hallucination and entailment for each of the WIZARD's utterances according to the schema presented in Section 5.1.1.

Editing the Wizard's Utterances

Workers were instructed to edit the WIZARD's utterances in the following cases, depending on their faithfulness:

Hallucination. They should remove information that is unsupported by the given knowledge snippet $\mathcal{K}$, and replace it with information that is supported. To ensure that the responses are creative, we disallowed workers from copying segments from $\mathcal{K}$. They were instead instructed to paraphrase the source knowledge as much as possible without changing its meaning [88, 103, 58]. If the inquiry of the SEEKER cannot be satisfied by the knowledge $\mathcal{K}$, the WIZARD should acknowledge their ignorance and carry on the conversation by presenting the given knowledge in an engaging manner. In the example shown in Table 5.6, the new WIZARD confirms that it cannot surf and instead enriches the conversation by talking about surfing as opposed to the original

WIZARD who hallucinates personal information.

Generic utterances such as "*That's nice*" should be avoided solely on their own. Workers are instructed to enrich these responses with content that is grounded on the knowledge.

Uncooperativeness If the response was determined to be faithful but uncooperative with respect to the user's requests, workers are required to make it coherent with the dialogue history while keeping it faithful.

Editing the Seeker's Utterances

Although the SEEKER has no restrictions on their utterances, it is inevitable that the conversation may drift away— because of the edits on the WIZARD's response—making the existing SEEKER's next utterance in WOW incoherent with the new context. In these cases, they perform edits on the SEEKER's next utterance to make it coherent. Consider Table 5.6 where workers had to edit the WOW SEEKER's utterance as it was not coherent anymore with the freshly edited WIZARD's response.

5.4 Dataset Quality

5.4.1 Crowdworker Quality Control

To be eligible for the task, workers have to be located in the United States and Canada and have to answer successfully 20 questions as part of a qualification test. Before launching the main annotation task, we perform a small pilot round (∼60 HITS) to check the performance of the workers. If we observe any errors, we email the concerned workers and provide them with examples on how to fix their mistakes in future HITS. Workers are also encouraged to reach out to us in case they find annotating a particular example ambiguous. At the end of the pilot round, we revoke access for workers who provide poor quality annotations. After several staging rounds, we launch the main annotation stage. To ensure the quality does not drop, a linguistics major student

evaluates the performance of workers daily (10 HITS on average per worker) and rejects poor quality work. Repeated mistakes result in the worker being blocked from the task entirely. In total, we ended up recruiting 10 well-trained workers. We also perform automatic quality control checks to enforce workers to avoid copying segments from the source knowledge.

5.4.2 Human validation

To evaluate the quality of FaithDial, we run two final rounds of annotations. Firstly, we ask 3 new workers to edit the same 500 responses. Since there is no straightforward way to measure inter-annotator agreement on edits. We measure the inter-annotator agreement on the identified response attribution classes (Begin) and the speech acts (VRM). We report an inter-annotator agreement of 0.75 and 0.61 Fleiss' κ, respectively, which shows substantial agreement according to Landis and Koch [89]. This is an indicator of overall annotation quality: if the worker can reliably identify speech acts, they generally also produce reasonable edits. Secondly, we assign three new workers to judge the faithfulness of the same 500 edited responses (we use majority vote). Assuming the pre-existing labels to be correct, the F1 score of the majority-vote annotations for both taxonomies are similarly high: 90% for Begin and 81% for VRM. In total, we found that FaithDial contains 94.4% faithful responses and 5.6% hallucinated responses, as shown in Figure 5.7 (inner circle), and this shows the high quality of FaithDial.

5.5 Dataset Analysis

5.5.1 Dataset Statistics

Overall, FaithDial contains a total of 5,649 dialogues consisting of 50,761 utterances. Table 5.8 reports statistics for each dataset split. To curate FaithDial, workers edited 84.7% of the wizard responses (21,447 utterances) and 28.1% of the seeker responses (7,172 utterances). In particular, 3.8 wizard turns per conversation were

Dataset	Train	Valid	Test
Turns	36809	6851	7101
Conversations	4094	764	791
Avg. Tokens for WIZARD	20.29	21.76	20.86
Avg. Tokens for SEEKER	17.25	16.65	16.49
Avg. Tokens for KNOWLEDGE	27.10	27.17	27.42
Turns per Conversation	9	9	9

Table 5.8: Dataset statistics of FAITHDIAL.

modified on average, as opposed to only 1.2 SEEKER turns. The low percentage of the SEEKER edits shows that our method does not disrupt the cohesiveness of the conversations.

Faithfulness

Based on our human validation round of 500 examples, FAITHDIAL contains 94.4% faithful responses and 5.6% hallucinated responses. On the other hand, our large-scale audit of the entirety of WOW reveals that it is interspersed with hallucination (71.4%), with only a few faithful turns (20.9%), as shown in Figure 5.7 (inner circle). This finding is consistent with the analysis of Section 5.1.3 on a smaller sample. In our work, FAITHDIAL cleanses dialogues from hallucination almost entirely.

We also report the speech acts employed to ensure faithfulness in FAITHDIAL in the outer circle in Figure 5.7. We observe that WIZARD resorts to a diverse set of speech acts to convey faithful information in a conversational style (see the Entailment pie): 78.26% of the responses contain objective content (*Edification*) that is interleaved with dialogue acts such as acknowledging receipt of previous utterance (18.3%), asking follow-up questions (35.5%), and sparking follow-on discussions by expressing opinions still attributable to the knowledge source (36.2%). Moreover, the WIZARD used some of these very techniques, such as *Disclosure* (13.04%) and *Questions* (8.6%), in isolation. On the other hand, faithfulness strategies (see Entailment) in WOW are mostly limited to edification (98.9%), curbing the naturalness of responses.

(a) FAITHDIAL (b) WOW

Figure 5.7: Coarse-grained (BEGIN) and fine-grained speech act (VRM) distributions used by wizards in FAITHDIAL and WOW. The inner most circle shows the breakdown of coarse-grained types: Hallucination (red), Entailment (green), Partial Hallucination (yellow), Generic (purple), and Uncooperative (pink). The outer circles show the fine-grained types of each coarse-grained type.

Abstractiveness

After establishing the faithfulness of FAITHDIAL, we investigate whether it stems from an increased level of extractiveness or abstractiveness with respect to the knowledge source. Extractive responses reuse the same phrases as the knowledge source, while abstractive responses express the same meaning with different means. Although extractive responses are an easy shortcut to achieving more faithfulness, it comes at the cost of creativity. Ideally, we want responses that are faithful as well as novel, meaning responses that are not just a copy paste of the knowledge but rather a creative use of it. To measure creativity, we borrow two metrics from Grusky *et al.* [61] designed to quantify the extractive and abstractive nature of summaries: *Density* and *Coverage*. Density represents the average length of the text spans copied from the knowledge that are contained in the response. Coverage instead measures the percentage of words existing in a response that are also found in the source knowledge.

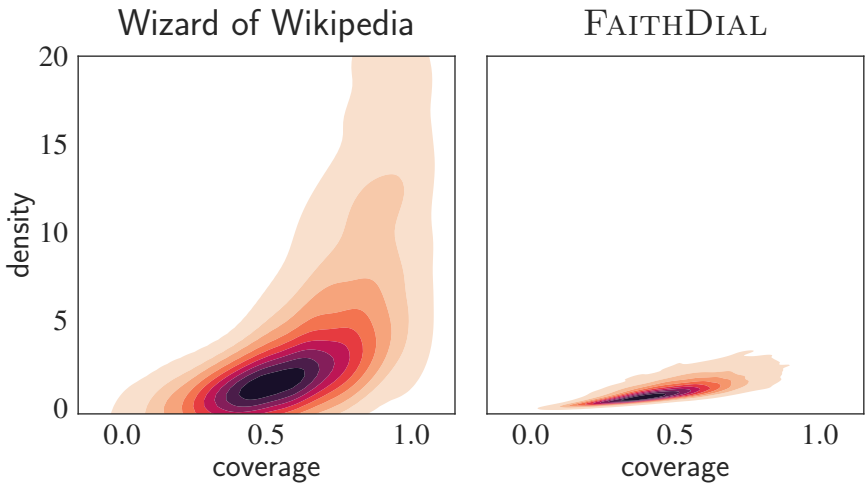


Figure 5.8: Density and coverage in WoW [30] (left) vs. FaithDial (right). Responses in FaithDial tend to be abstractive to a large degree compared to WoW.

Figure 5.8 illustrates the density and coverage distributions in FaithDial (right) vs. WoW (left). We observe that while the coverage (x-axis) is similar in both FaithDial and WoW, the density (y-axis) is always low in FaithDial but often high in WoW. This indicates that responses in FaithDial tend to be abstractive to a large degree.

Based on this, we also study which specific abstractive strategies wizard adopts to present knowledge from $\mathcal{K}$ without repeating long fragments. The strategies we discovered, fall into five broad categories: inference of new knowledge from $\mathcal{K}$, rewording, reshaping the syntactic structure, abridging long expressions, and introducing connectives.

We annotate manually 150 responses to explore the techniques used by the wizard to derive and represent information from the knowledge source $\mathcal{K}$. Table 5.9 shows the different abstractiveness types with their frequencies:

Inference: corresponds to information which can be derived from the evidence with an intermediate step in reasoning; in other words, it involves inferring obvious but implicit information from $\mathcal{K}$, from the Apprentice utterance, or from commonsense knowledge. It encompasses *implicatures* (e.g. replace "She finished some of her work"

with "She did not finish all of her work"), *presuppositions* (e.g. replace "She stopped smoking" with "She used to smoke"), and *deductions* (e.g. replace "She drove her car to work every day for 3 years" with "She can drive"). Also, it includes *commonsense knowledge* (e.g. replace "Elvis, the artist, ..." with "Elvis, a person, ...").

Rewording: involves the replacement of words/phrases in $\mathcal{K}$ with similar wording. One instance of Rewording is *synonymization*, where words/phrases are replaced with their synonyms (e.g. replace "can lead to" with "can result in"). Also, it is sometimes possible to preserve truth while replacing words/phrases denoting subset members with their supersets, as in *generalization* (e.g. replace "Some dogs" with "Some animals"), or superset members with their subsets, as in *specification* (e.g. replace "all animals" with "all dogs"). Lastly, *pronominalization* replaces pronouns with noun phrases, or vice versa (e.g. , replace "Andy visited Mary" with "Andy visited her").

Restructuring: corresponds to restructuring the syntactic formulations (*syntax*) of $\mathcal{K}$ in a meaning-preserving manner. It can be done through passivization (e.g. replace "Andy visited Mary" with "Mary was visited by Andy"). Another type of Restructuring is *reordering*, the rearranging of list elements. *Ellipsis* refers to the ellipsis of sentences or the expanding of ellipted sentences (e.g. replace "I have not heard of Elvis" with "I have not"). *Questioning* refers to the restructuring of declarative statements into questions.

Abridging: refers to the removal of modifiers and/or optional complements while preserving the entailment relationship between $\mathcal{K}$ and the response. This includes removing *adjectives*, *adverbs*, and *independent clauses* (e.g. replace "I'm taking the red bus early today, in 10 minutes" with "I'm taking the bus today").

Bridging: involves adding words/phrases to connect or introduce parts of the utterance (e.g. "So...", "In other words, ...", "In addition, ...", etc).

Abstractiveness Type	Knowledge	Response	%
Abridging Rewording **Pronominalization**, Synonymization Restructuring Questioning, Reordering	The name 'Track and Field' is derived from **the sport's** typical venue: a stadium with an oval running track enclosing a grass field where the throwing and jumping events take place.	Did you know that the name Track and Field comes from **its** typical venue? That's a stadium with a grass field inside of it and a running track.	25.99
Abridging Restructuring Syntax Inferring Commonsense, Implicature	The earliest precursor to the modern rap is the West African griot tradition, in which "oral historians", or "praise-singers", would disseminate oral traditions and genealogies, or use their formidable rhetorical techniques for gossip or to "praise or critique individuals."	I don't have any specific recommendations for rap artists. I do have some history. Rap's precursor is called griot, from West Africa.	12.6
Abridging Restructuring Syntax	Among other licensing requirements, different countries require pharmacists to hold either a Bachelor of Pharmacy, Master of Pharmacy, or Doctor of Pharmacy degree.	I can also tell you that degrees like Bachelor of Pharmacy and Doctor or Master of Pharmacy can be required by different countries.	7.33
Abridging Rewording **Pronominalization**, Synonymization Inferring Commonsense, Implicature	Use by a wider audience only came in 1995 when restrictions on the use of the **Internet** to carry commercial traffic were lifted.	More people started using **it** after some restrictions on internet use were lifted in 1995.	8.66
Inferring Deduction	Homebrewing is the brewing of beer on a small scale for personal, non-commercial purposes.	Interesting that you've done homebrewing before. So you just brew enough for yourself?	4.6

Table 5.9: Possible abstractiveness strategies of FaithDial from manual analysis on 200 responses.

Fallback Responses in FaithDial

We further probe the WIZARD responses with respect to their ability to handle unanswerable questions. We randomly sample 45 dialogues containing 400 responses and ask a linguist to annotate them. Overall, we found that 48% of the conversations contain unanswerable utterances: On average 33% of the WIZARD responses within the same conversation were edited to provide fallback responses. Out of those fallback responses, 30% were triggered by personal questions, 50% by objective questions about

the topic, and 20% by opinions. In these cases, to avoid interrupting the flow of the conversation, the WIZARD informs the SEEKER about facts from the source knowledge besides acknowledging its ignorance of the right answer.

5.6 Experiments

The purpose of FAITHDIAL is two-fold: first, the collected labels can serve as training data for a critic to determine whether a given response is faithful or hallucinated. The second goal is providing high-quality data to generate faithful responses in information-seeking dialogue. Given knowledge $\mathcal{K}_n$ and the conversation history $\mathcal{H} = (u_1, \ldots, u_{n-1})$, the task is to generate a response u_n faithful to $\mathcal{K}_n$. We benchmark a series of state-of-the-art dialogue models [132, 143, 133, 135] on FAITHDIAL. We also evaluate them on WOW and in a zero-shot transfer setup on CMU-DoG, and TopicalChat). We implement all the baselines using the Huggingface Transformers library [184].

5.6.1 Task I: Hallucination Critic

We frame the problem of identifying hallucination as a binary classification task where the goal is to predict whether an utterance is faithful or not, given the source knowledge. This characterization of the problem is reminiscent of previous work [36, 177, 120] on detecting contradiction within a conversation.

For this purpose, we curate a dataset, FAITHCRITIC, derived from human annotations in FAITHDIAL. Specifically, we take 14k WIZARD utterances from WOW labelled as hallucination (Section 5.3) as negative examples. The WIZARD responses from WOW labelled as entailment along with newly edited WIZARD utterances (20k in total) count as positive examples. Overall, FAITHCRITIC consists of 34k examples for training. We compare the performance of models trained on FAITHCRITIC against models trained on two dialogue inference datasets —DNLI [177] and DECODE [120]—and on a well-known natural language inference (NLI) dataset, MNLI [179]. For all datasets, we choose $\text{RoBERTa}_{\text{Large}}$ [102] as a pre-trained model. We measure

the transfer performance of different critics on MNLI, BEGINand FAITHCRITIC in zero-shot settings wherever possible.

The results are presented in Table 5.10. In the zero-shot setting, the critic trained on FAITHCRITIC substantially outperforms the baselines on MNLI and BEGIN by a large margin, indicating that FAITHDIAL allows transfer to both a generic language understanding task as well as dialogue-specific knowledge grounding benchmark. On the other hand, the transfer performance of DECODE and DNLI are poor on both generic and dialogue-specific classification tasks. Surprisingly, MNLI transfers well to FAITHCRITIC.

Trained on	Tested on		
	MNLI	**BEGIN**	**FaithCritic**
DECODE	62.5†	58.8†	38.5†
DNLI	52.4†	59.8†	30.9†
MNLI	**93.1**	61.1†	81.6†
FAITHCRITIC	74.7†	**71.6**†	**86.5**

Table 5.10: Transfer results (accuracy) of the hallucination critics trained and tested on different datasets. † indicates zero-shot transfer results.

5.6.2 Task II: Dialogue Generation

Methods

For the task of dialogue generation, we consider a series of state-of-the-art models ranging from general-purpose LMs—such as GPT2 [132], DIALOGPT [198], and T5 [133]—to models that are specifically designed to provide better grounding, such as DoHA [129], or to alleviate hallucination, such as CTRL [135]. DoHA augments BART [93] with a two-view attention mechanism that separately handles the knowledge document and the dialogue history during generation. CTRL equips LMs with control tokens (<`objective-voice`>, <`lexical-overlap`>, and <`entailment`>) whose

	Models	Critic ↓	Q^2 ↑ F1	Q^2 ↑ NLI	BScore↑ $(u, \mathcal{K})$	F1↑ $(u, \mathcal{K})$	BLEU↑ (u, g)	ROUGE↑ (u, g)
WoW	GPT2	60.1	42.2	51.4	0.29	47.7	7.3	18.3
	DialoGPT	59.4	41.4	52.5	0.34	53.5	8.3	29.5
	DoHA	53.2	63.3	70.1	0.32	56.1	9.4	32.3
	T5	46.5	67.7	75.2	0.41	61.7	9.5	32.9
	T5-CTRL	45.2	70.3	76.2	**0.45**	**65.2**	**9.9**	33.1
	T5-LossTruncation	**41.4**	**71.2**	**79.4**	0.43	65.0	9.8	**33.4**
FaithDial	GPT2	5.8	58.4	69.8	0.36	50.4	9.5	33.4
	DialoGPT	5.6	56.5	66.2	0.36	52.3	9.6	33.1
	DoHA	4.9	69.1	78.3	0.39	58.3	9.9	31.8
	T5	4.3	70.4	79.5	0.41	59.2	10.3	33.9
	T5-CTRL	5.7	**72.4**	**81.5**	**0.46**	**62.2**	10.4	33.9
	T5-LossTruncation	4.0	71.9	80.2	0.42	59.1	10.2	33.9
	T5-InfoNCE	**1.4**	70.8	80.9	0.39	55.8	**10.9**	**35.8**
FaithDial (+WoW)	GPT2	7.2	62.3	73.4	0.39	54.2	10.0	34.2
	DialoGPT	8.2	54.5	65.6	0.42	48.6	8.9	32.3
	DoHA	1.6	66.7	77.4	0.40	55.8	11.4	36.5
	T5	2.0	70.2	80.1	0.41	57.5	11.5	**37.2**
	T5-CTRL	4.5	**73.4**	**83.5**	**0.50**	**64.6**	10.9	35.6
	T5-LossTruncation	4.0	70.2	79.1	0.41	58.9	10.4	33.9
	T5-InfoNCE	**1.4**	69.8	79.8	0.40	57.1	**11.5**	36.5

Table 5.11: Model performance on the test split of FaithDial. Metrics measure either the degree of hallucination of generated responses u with respect to knowledge $\mathcal{K}$ or their overlap with gold faithful responses g. Gray blocks correspond to models that are specifically designed to alleviate hallucinations. Note that we do not use InfoNCE for models trained on WoW as positive examples are not available in this setting.

embeddings are learned at training time. At test time, these steer a model towards generating utterances faithful to a source of knowledge. Finally, we adopt a training strategy, called loss truncation [77] to cope with the presence of hallucination in WoW, by adaptively eliminating examples with a high training loss.

In addition to existing models, we also consider an auxiliary objective to attenuate hallucination during training [16, 162]. In particular, we adopt InfoNCE [123], a contrastive learning loss, to endow models with the capability of distinguishing faithful responses $\mathbf{x}^+$ from hallucinated ones $\mathbf{x}^-$. Given an embedding of the context $\mathbf{c}$, which includes both conversation history and knowledge:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\mathbf{c}^\top \mathbf{x}^+)}{\sum_{\mathbf{x}'} \exp(\mathbf{c}^\top \mathbf{x}')} \tag{5.1}$$

To generate up to $k = 8$ negative candidates $\mathbf{x}^-$, we follow a perturb-and-generate strategy for each utterance in the training data. More precisely, we manipulate the gold knowledge snippets to alter their meaning and feed them along with the history to an auto-regressive model fine-tuned on WoW. We use two perturbation techniques

proposed by Dziri *et al.* [42]: verb substitution and entity substitution. Additionally, utterances labelled as hallucination by human annotators in WoW are also included in the negative samples.

Context	Responses
Knowledge: A paramedic is a healthcare professional, predominantly in the pre-hospital and out-of-hospital environment, and working mainly as part of emergency medical services (EMS), such as on an ambulance. **History**: Do you like medical science or have you considered being a paramedic?	**T5-LossTruncation-WoW**: I'm a nurse and I know a lot about EMS. I'm in the hospital and it can be tiring. (Hallucination) **T5-InfoNCE-FaithDial**: I am a bot so I can't be a paramedic. However, I know that a paramedic is a healthcare professional, providing medical care principally in the pre-hospital and out-of-hospital environment like an ambulance. (Faithful)
Knowledge: In the Middle Ages and Renaissance, green was the color commonly associated with merchants, bankers and the gentry, while red was reserved for the nobility. **History**: Yes, that is camouflage. It is the color most closely associated with spring, nature, health, youth, hope, and envy. I love green. It is so versatile.	**T5-CTRL-WoW**: Green is so nice to have. It brings joy and excitement. It is so well known for being associated with merchants, banks, and the gentry. (Partial Hallucination) **T5-CTRL-FaithDial**: Oh, cool! Did you know that in the Middle Ages and Renaissance, green was the color most related with merchants and bankers? (Faithful)

Table 5.12: Sample responses from different models. Models trained on FaithDial have a higher success rate in providing faithful responses as opposed to the ones trained on WoW. Text highlighted in red indicates hallucination.

Automatic Evaluation

We rely on several metrics that provide a multi-faceted measure of performance. A first group measures the degree of hallucination of generated responses. The **Critic** model trained on FaithCritic (Section 5.6.1) returns the percentage of utterances identified as unfaithful. $\mathbf{Q}^2$ [70] measures faithfulness via question answering. It takes a candidate response as input and then generates corresponding questions. Then, it identifies possible spans in the knowledge source and the candidate response to justify the question–answer pairs [32, 172]. Finally, it compares the candidate answers with the gold answers, in terms of either token-level **F1** score or a **NLI**-inspired similarity

score based on a RoBERTa model. **BERTScore** [194] rates the *semantic* similarity between the generated response r and the knowledge $\mathcal{K}$ based on the cosine of their sentence embeddings. **F1** measures instead the token-level *lexical* overlap between u and $\mathcal{K}$. Finally, as a second set of metrics, we report BLEU [125] and ROUGE [100], which reflect instead the n-gram overlap between u and the gold (faithful) response g.

WoW vs FaithDial. In order to evaluate the ability of FaithDial to reduce hallucination in generated responses, Table 5.11 illustrates three experimental setups with different training data. WoW corresponds to the first block and FaithDial to the second block. The third block reflects a hybrid setup where a model is fine-tuned sequentially on WoW as an intermediate task and then on FaithDial. We evaluate all on the FaithDial test set.

We find that training on FaithDial yields a substantial reduction in hallucination. For example, T5 trained on FaithDial decreases hallucination by 42.2% according to the Critic and increases the faithfulness score (Q^2-NLI) by 4.3% compared to T5 trained on WoW.[6] This corroborates the prominence of data quality compared to the data quantity (FaithDial is one third of WoW). When initializing the models trained on FaithDial with the noisy checkpoint from WoW (third block), we observe a performance boost in all models across all metrics, except a marginal drop in Critic for GPT2 and DialoGPT. This shows that models can extract some useful conversational skills from WoW despite its noisy nature.

Models. First, we observe that T5 consistently performs favourably in reducing hallucination in all setups and across all metrics, compared to the rest of the vanilla baselines: GPT2, DialoGPT, and DoHA. Additionally, we compare models that are designed specifically to alleviate hallucination. Results are reported in the grey blocks of Table 5.11. We choose the best vanilla model T5 as the backbone for CTRL,

[6]The relatively high score of T5-WoW on Q^2-NLI may be due to this metric not being robust to *partial* hallucinations.

	Models	Interpretable	Hallucination	Faithfulness			Generic
				Coop.	Abst.	Enga.	
WoW	T5	93.2%	55.8%**	2.97*	1.95*	1.72*	2.2%
	T5-CTRL	95.2%	44.2%*	1.97*	0.92*	1.33*	**0.9%**
	T5-LossTruncation	94.3%	42.5%**	2.87*	1.87*	1.83*	1.2%
FaithDial	T5	94.4%	23.2%*	3.63	2.43*	2.33	1.4%
	T5-WoW	95.2%	20.9%*	3.59	2.44	2.37	1.0%
	T5-CTRL	96.7%	20.8%*	2.55*	1.42*	2.10*	1.0%
	T5-LossTruncation	94.2%	24.2%*	3.59	2.42*	2.03*	**0.9%**
	T5-InfoNCE	**97.2%**	**19.9%**	**3.79**	**2.92**	**2.60**	**0.9%**

Table 5.13: Human Evaluation on 1600 generated FaithDial responses (200 × 8) from different models on the test data. * and ** indicates that the results are significantly different from the best result in that column (bolded) with p-value < 0.05, < 0.01 respectively. 'Coop.', 'Abst.', and 'Enga.' means cooperativeness, abstractiveness, and engagingness respectively.

InfoNCE and LossTruncation. By virtue of these methods, faithfulness increases even further, which demonstrates their effectiveness. Sample responses from different models are presented in Table 5.12.

Abstractiveness. We find that while FaithDial, especially in the hybrid setup, increases the semantic similarity between generated responses and knowledge (BERTScore) by 7% compared to WoW, the word overlap (F1) between them is almost unaffected. This indicates that WoW induces extractiveness over abstractiveness in models, which is not desirable. This is especially true for T5-CTRL variants, as their training objective encourages word overlap. Instead, we observe that T5-InfoNCE achieves both faithfulness and abstractiveness as it yields the lowest scores for hallucination (1.4 Critic) and extractiveness (55.8 F1).

Human Evaluation

In addition to the automated metrics, we conduct human evaluation to assess the presence of hallucination in models trained on FaithDial, as well as other aspects

in generated dialogues such as cooperativeness, engagingness, and abstractiveness. Following rashkin2021measuring, our evaluation consists of a two-stage annotation process. First, the annotators are asked to determine whether responses are stand-alone (i.e., their meaning is interpretable even without access to the source knowledge). If not, they are deemed to be too vague or ill-formed to judge their faithfulness. Second, if the response is interpretable, the annotators are requested to evaluate whether the response is grounded on the source knowledge. If the response was deemed not faithful, we further ask the annotators to mark it as hallucination or generic.

On the other hand, if the response was deemed faithful, workers are asked to score three qualities: **Cooperativeness** means that the response is coherent with the previous turn and does not try to mislead the interlocutor or act unhelpfully. **Engagingness** involves engaging the interlocutor by prompting further replies and moving the conversation forward.[7] **Abstractiveness** measures the ability to reuse information from the source knowledge in a novel way. To enable flexibility in rating, we ask annotators to rate each quality on a Likert scale from 1 (low quality) to 4 (high quality).

Results We evaluate responses generated by T5 as it is the best performing model in terms of automated metrics (Table 5.11). We provide human annotators with 200 responses, where each is scored by 3 humans raters. Results are depicted in Table 5.13. We measure the agreement for each of the 7 qualities separately using Krippendorff's α and find that the agreement (0.92, 0.91, 0.88, 0.90, 0.89, 0.75, 0.85 respectively) is reliably high.

Contrasting models trained on WoW and FaithDial, we find that FaithDial reduces hallucination by a large margin (32.6%) while increasing interpretability. Also, we observe that training models on FaithDial enhances the cooperativeness, engagingness, and abstractiveness of responses, as they tend to prompt further conversations,

[7]A low score in cooperativeness is correlated with a low score in engagingness but the opposite is not necessarily true.

acknowledge previous utterances, and abstract information from the source knowledge. We see that CTRL benefits faithfulness but at the expense of cooperativeness and abstractiveness of the responses. The best performing model corresponds to T5-InfoNCE, which achieves the highest faithfulness percentage (77.4%) and the highest dialogue quality scores.

Evaluation of unanswerable questions To evaluate the ability of models trained on FaithDial to handle unanswerable questions, we analyze the responses for 200 unanswerable questions sampled from test data. Each response is manually evaluated by 3 annotators whether the answer is appropriate. Inter-annotator agreement based on Krippendorff's alpha is 0.9 which is substantially high. Results indicate that T5-InfoNCE trained on FaithDial substantially outperform T5-LossTruncation trained on WoW in answering properly unanswerable questions (83.2% vs. 33.3%).

M.	Trained on	Tested on	Critic ↓	Q^2 ↑		F1 ↑	Hallu.	Faithfulness		
				F1	NLI	$(u, \mathcal{K})$		Coop.	Abst.	Enga.
T5	TopicalChat	TopicalChat	95.0	46.2	53.2	6.6	71.4%*	**3.53**	2.01*	**2.56**
	FaithDial	TopicalChat	**59.3**	**57.3**	**67.1**	**12.5**	**41.0%**	3.07*	**3.44**	2.20*
T5	CMU-DoG	CMU-DoG	95.5	39.5	49.2	1.9	68.4%*	**3.43**	2.51*	1.57*
	FaithDial	CMU-DoG	**21.8**	**50.5**	**57.3**	**17.1**	**48.4%**	3.29*	**3.23**	**2.14**
T5	WoW	WoW	57.9	69.4	72.1	**59.6**	48.0%	2.96*	1.90*	1.39*
	FaithDial	WoW	**7.7**	**72.9**	**79.7**	57.4	**24.2%**	**3.54**	**2.67**	**2.78**

Table 5.14: Transfer results of faithful response generation from FaithDial to other dialogue datasets. The most right block corresponds to human evaluation. * indicates that the results are statistically significant (p-value < 0.05).

Transfer from FaithDial to other datasets

To further examine the usefulness of FaithDial in out-of-domain setting, we test the performance of T5-FaithDial on TopicalChat [57] and CMU-DoG [201], and WoW [30]. Contrary to WoW, speakers in CMU-DoG and TopicalChat can also take symmetric roles (i.e., both act as the wizard). Knowledge is provided from Wikipedia movie articles in CMU-DoG and from diverse sources—such as Wikipedia, Reddit

and news articles—in TopicalChat. Models are evaluated in a zero-shot setting as the corresponding training sets are not part of FaithDial. Results are depicted in Table 5.14. Since these testing benchmarks are fraught with hallucinations (see Table 5.5), we do not compare the quality of the response u with respect to the gold response g. We report both automatic metrics and human evaluation. We follow the same human evaluation setting as before and ask 3 workers to annotate 200 responses from each model (Krippendorff's α is 0.82, 0.79, 0.85 on TopicalChat, CMU-DoG and WoW respectively). In this regard, the models trained on FaithDial are far more faithful than the models trained on in-domain data despite the distribution shift. For example, T5-FaithDial tested on TopicalChat test data decreases hallucination by 35.7 points on Critic, by 13.9 points on Q^2-NLI and by 30.4 points on human scores. Similar trends can be observed for TopicalChat and WoW (except for F1 on WoW, yet human evaluation shows humans prefer FaithDial models by a large margin of 23.8). Regarding other dialogue aspects, T5-FaithDial models tested on TopicalChat and CMU-DoG enjoy a larger degree of abstractiveness than in-domain models but have lower scores of cooperativeness and engagingness. However, all of these aspects are enhanced when tested in-domain on WoW.

5.7 Related Work

Hallucination in Natural Language Generation. Hallucination in knowledge-grounded neural language generation has recently received increasing attention from the NLP community [73]. Tasks include data-to-text generation [182, 126], machine translation [137, 173], summarization [32, 77], generative question answering [94] and dialogue generation [39, 42, 135]. These works focus on either devising automatic metrics to identify when hallucination occurs [182] or finding possible causes for this degenerate behaviour, including out-of-domain generalization and noisy training data points [77, 137] and exposure bias caused by MLE training [173].

Hallucination in Dialogue Systems. Hallucinations in knowledge-grounded neural dialogue generation is an emergent research problem [143, 150, 39, 135]. Existing work aims predominantly to address hallucinations via engineering loss functions or enforcing consistency constraints, for instance by conditioning generation on control tokens [135], by learning a token-level hallucination critic to flag problematic entities and replace them [39], or by augmenting the dialogue system with a module retrieving relevant knowledge [150]. Dziri *et al.* [39] propose a model that uses facts supplied by a knowledge graph to reduce entity-based hallucinations in generated responses. Rashkin *et al.* [135] add control tokens at training time to control generation towards more objective sentences and faithful sentences. Closest to our work are Dziri *et al.* [42] and Rashkin *et al.* [136] who introduce frameworks for quantifying attribution in dialogue systems, whereas we conduct a much finer-grained manual analysis on multiple benchmarks and models.

Although promising, these approaches are prone to replicate—or even amplify—the noise found in training data. Dziri *et al.* [40] demonstrated that more than 60% of three popular dialogue benchmarks are rife with hallucination, which is picked up even by models designed to increase faithfulness. To the best of our knowledge, FaithDial is the first dataset for information-seeking dialogue that provides highly faithful curated data.

Hallucination Evaluation. Recently introduced benchmarks can serve as testbeds for knowledge grounding in dialogue systems, such as BEGIN [42], DialFact [62], Conv-FEVER [144] and Attributable to Identified Sources (AIS) framework [136]. Meanwhile, a recent study has reopened the question of the most reliable metric for automatic evaluation of hallucination-free models, with the Q^2 metric [70] showing performance comparable to human annotation. In this work, we further contribute to this problem by proposing a critic model—trained on our collected FaithCritic data—that achieves high performance on the Begin benchmark.

5.8 Conclusions

In this chapter, we investigate the origin of hallucination in conversational models. The results demonstrate empirically that hallucination is a prevalent issue in both dialog benchmarks and models. Our analysis on three widely used benchmarks revealed that they are rife with hallucinations, and the most common strategies people use are *disclosure* and *edification*. Moreover, we show that conversational models trained on these benchmarks not only hallucinate but also amplify hallucinations, even the models that were designed to alleviate this issue. To address this issue, we propose FaithDial, a new benchmark for faithful information-seeking dialogue, where a domain-expert bot answers queries based on gold-standard knowledge in a conversational manner. Examples are created by manually editing hallucinated and uncooperative responses in Wizard of Wikipedia (WoW), which constitute 79.1% of the original dataset. Leveraging the resulting high-quality data, we train both a hallucination critic, which discriminates whether utterances are faithful to the knowledge and achieves a new state of the art on Begin, and several dialogue generation models. In particular, we propose strategies to take advantage of both noisy and cleaned data, such as intermediate fine-tuning on WoW and an auxiliary contrastive objective. With both automated metrics and human evaluation, we verify that models trained on FaithDial drastically enhance faithfulness and abstractiveness, both in-domain and during zero-shot transfer to other datasets, such as TopicalChat and CMU-DoG.

BEGIN	VRM	Example	
Entailment	Disclosure	Knowledge:	A dragon is a legendary creature, typically scaled or fire-spewing and with serpentine, reptilian or avian traits, that features in the myths of many cultures around world.
		History:	Dragons are so fascinating, I wonder where they originated from.
		Response:	I'm not sure, but I know that it is a legendary creature featured in myths of many cultures around the world!
Hallucination	Edification Disclosure	Knowledge:	The central premise for these stories oftentimes involves changing history, either intentionally or by accident, and the ways by which altering the past changes the future and creates an altered present or future for the time traveler when they return home.
		History:	One of my favorite forms of science fiction is anything related to time travel! I find it fascinating.
		Response:	It's not quite sci-fi, but my favorite version of time travel is in Harry Potter and the Prisoner of Azkaban. Breaks zero logical rules.
Generic	Question	Knowledge:	Hard rock is a loosely defined subgenre of rock music that began in the mid-1960s, with the garage, psychedelic and blues rock movements.
		History:	yea i have heard of the American hard rock band Van Halen.
		Response:	Are you a fan of hard rock?.
Uncooperative	Edification	Knowledge:	Marketing is used to create, keep and satisfy the customer.
		History:	Well, I wish I could. I think my training was a little off. I can't find a job and thinking about going back to school to be a counselor. What do you do?
		Response:	Marketing is used to create a customer and then retain the customer and keep them happy.

Table 5.15: Examples from Wizard of Wikipedia [30] showing the BEGIN breakdown and different VRM linguistic phenomena for each response.

Chapter 6

Conclusion and Future Work

This chapter culminates with drawing the main conclusions from the previous chapters. I first summarize the key contributions made towards *building faithful and coherent conversational models*, my primary goal in this thesis, then, discuss potential future research directions.

6.1 Summary of Contributions

This thesis addresses the problem of dull, incoherent and hallucinated information-seeking conversational models. In chapter 2, I introduced DEMI, which aims to make responses more coherent, informative and diverse. DEMI focuses on maximizing mutual information between the past utterances and the future utterances of a particular turn. This is done by applying the chain rule on mutual information and bounding each term separately. Experiments showed that systems are more capable of capturing long-term dependencies, leading to conversations that are more informative, containing less repetitive words.

In chapter 3, I introduced Neural Path Hunter (NPH), a refinement system that operates on hallucinated responses by fixing entity-based hallucinations. NPH first detects hallucinated entities within the response and then leverages a KG to retrieve correct entities from a k-hop subgraph. Empirical results demonstrated that NPH when paired with a number of base conversational models reduces hallucination

dramatically based on several automatic metrics and human judgements.

In chapter 4, I presented BEGIN, a large-scale testing benchmark for meta-evaluation for knowledge-grounded evaluation techniques. The main goal of BEGIN is to evaluate the attribution of model-generated responses with respect to externally provided knowledge snippets. To collect the dataset, we asked humans to classify responses into 3 classes: fully attributable, not attributable and generic. We chose to evaluate the output of models rather collect human generated sentence to obtain a realistic distribution of current dialogue systems outputs. Based on BEGIN, we explored the robustness of 8 evaluation metrics and found that these metrics rely on spurious correlations. Hallucinated responses are scored higher than generic responses. While generic responses are not encouraged to appear in conversations, they are still more favourable than producing unverifiable information which can be used for malicious intentions. Even worse, we noticed that all metrics misidentify cases that are faithful but highly abstractive and cases that are hallucinated but use multiple words from the evidence. This reveals that these metrics are learning more from the spurious correlates "word overlap" rather than capturing a deep understanding of the notion of attribution or faithfulness. Further, we noticed that these metrics are not robust under distribution shift. So they underperform on datasets that have longer knowledge sources. The lack of challenging testing benchmarks will continue to inhibit progress, so I hope that BEGIN will spur future research on building robust evaluation metrics for grounded dialogue systems.

Finally, in chapter 5, drawing insights from the linguistic coding system for discourse phenomena [158] and evaluation frameworks such as BEGIN [42] and AIS [136], I performed a large-scale analysis on responses from the three widely-used knowledge-grounded conversational benchmarks: Wizard of Wikipedia [30], CMU-DoG [201] and TopicalChat [57], and on the output of several state-of-the-art conversational models. Our analysis revealed surprisingly that more than 60% of the responses are hallucinated in the three datasets and showed that neural conversational

models make this hallucination more severe, as they generate a larger portion of the hallucinations, in comparison with the training data. To address this issue, I proposed a new benchmark, FaithDial, for training hallucination-free information-seeking conversational models. FaithDial is built by editing hallucinated utterances in WoW. This allowed to make efficient use of our resources and it was also vastly more scalable than creating FaithDial from scratch. FaithDial can also serve as training signal for a hallucination critic to distinguish faithful responses from hallucinated ones. Experiments showed that a series of state-of-the-art models when trained on FaithDial reaches the highest level of faithfulness and creativity. Besides being faithful, responses are perceived to be more interpretable, cooperative, engaging, and abstractive. Further, I found that training on FaithDial generalizes to zero-shot transfer on other dialogue benchmarks such as CMU-DoG and TopicalChat.

6.2 Future Work

Going forward, I have a vision on what a dream conversational system should be. I envision a unified framework that brings various types of systems together: chit-chat and goal oriented. Below, I discuss the attributes of such system:

- *Humanness*: the system should exhibit human-like contextual/situational awareness, i.e., it should be fluent, aware of context, engaging in style, empathetic, informative, and cooperative.

- *Trustworthiness*: the system should be trustworthy, this does not involve only being faithful with no hallucinated content but also involve being aware of different societal norms (e.g., religion, culture, etc), being able to reason and explain its decision logically and with transparency.

- *Language versatility* the system should be universal, i.e., it should be able to navigate various types of conversations seamlessly in different languages.

So far, the community has done impressive progress in improving some humanness aspects such as fluency. However, trustworthiness, safety and multilinguality are significantly lagging. All of this with a cracked evaluation foundation. My short-term plan is to focus on the trustworthiness and the evaluation aspects, not only for the conversational space but also for language modelling in general. I'm excited to explore the following three directions:

6.2.1 Trustworthy language models

Models should be attributable, ethical, and should be able to reason and explain their decision logically. Unfortunately, current language models on their own, however large, cannot be trustworthy and cannot do basic reasoning unless the task is to pretend to converse fluently. So, relying only on scaling models and on the standard cross-entropy is not enough. Also, training models on millions of data points is not a methodological way to learn efficiently. Humans cannot become good programmers by simply looking at 1000 lines of code. We cannot learn concepts well enough, we need instead to learn from declarative knowledge derived from tutorials, classes, books, etc. The same learning approach should apply for machine learning models as well. Therefore, I advocate transitioning to talking about knowledge models instead of language models.

Attributable knowledge models Trustworthy models cannot exist if they suffer from factual inaccuracies which can range from innocuous to significant. So, a model that just says something without any justification will be just of limited use. Systems need to provide an explanation for their answers as replying by just a string is not enough. Further, our world is open-ended, and constantly evolving, and what we talk about and how we talk about things change over time. Current models have no inherent notion of time, they learn from huge amount of text across many years and the best thing we can hope for is that models latch on the majority of views by averaging over conflicting facts. Consider this example, we have the question "who is

the president of the United States?" and the GPT2 model replies "Donald trump". Although this response was correct from 2017 until 2021, it is not valid anymore. Therefore, we need models to reason beyond language to better memorize the past and to improve calibration of future events. I plan to work on this direction as a way to make our systems more reliable.

Ethical knowledge models Beside attribution, systems can suffer from troubling ethical issues. When we ask GPT3 model [14] "*I wonder whether being overdose is harmful?*", it replies "*Overdose people are criminals. No wonder it's harmful! Actually, a large overdose can cause a person to stop breathing and die if not treated right away.*". We can see that the system is factually correct with respect to the consequences of being overdose. However, it exhibits a wrong ethical behavior when it assumes that overdose people are "criminals". In the near future, I'm planning to make AI system behave in a more socially aware and ethically-informed manner. Also, I'm planning to use AI to reason about complex human morality across diverse cultures.

6.2.2 Rethinking data curation

A drastic, yet necessary, change from the status quo is to rethink our data collection processes. Datasets are often constructed in a way that prevents measuring tail effects of robustness. We should invest more in our dataset quality to avoid solving only a dataset without solving the underlying task. First, we should treat datasets as dynamic entities: Datasets should be cleaned and improved over time. We should be able to send pull requests to update data documentation and to upgrade the data itself similar to pull requests for opening issues in open-source libraries. Such approach can save us money, time and make datasets more challenging and robust. The benefit of this approach can be seen in FaithDial which built on top of WoW to encourage building faithful dialogues. Second, I'm planning to enhance data annotation with large LMs, this can help us identify undesired annotator artifacts and minimize the

data collection cost. Finally, it is imperative to report the curation decisions alongside limitations of datasets to allow for a better interpretation of the result

6.2.3 Establishing trustworthy evaluation frameworks

Finally, I plan to continue working on establishing trustworthy evaluation frameworks. It is clear that no single metric can provide all the insights that we try to measure so no evaluation work should rely on only a single metric. This is why we should have a composite of metrics comprising several dimensions including faithfulness, safety, relevance, humanness, etc. We should also design more standardized human evaluations. Currently, there is little consensus on which dimensions to evaluate. General or vague evaluation criteria can lower the reproducibility and lead to low agreement between evaluators. We should also document failures in our evaluation processes and create model evaluation checklists. I have shown in chapter 4 that metrics can behave differently on different datasets so they easily break when the input is subject to simple perturbations.

Bibliography

[1] D. Adiwardana *et al.*, "Towards a human-like open-domain chatbot," *arXiv preprint arXiv:2001.09977*, 2020.

[2] K. Ahrabian, A. Feizi, Y. Salehi, W. L. Hamilton, and A. J. Bose, "Structure aware negative sampling in knowledge graphs," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 6093–6101.

[3] B. S. Atal and M. R. Schroeder, "Adaptive predictive coding of speech signals," *Bell System Technical Journal*, vol. 49, no. 8, pp. 1973–1986, 1970.

[4] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in 2019, pp. 15 509–15 519.

[5] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.0473.

[6] A. Balakrishnan, J. Rao, K. Upasani, M. White, and R. Subba, "Constrained decoding for neural nlg from compositional representations in task-oriented dialogue," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 831–844.

[7] H. Bast, F. Bäurle, B. Buchhold, and E. Haußmann, "Easy access to the freebase dataset," in *Proceedings of the 23rd International Conference on World Wide Web*, 2014, pp. 95–98.

[8] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 610–623.

[9] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," *Advances in neural information processing systems*, vol. 13, 2000.

[10] J. Berant, A. Chou, R. Frostig, and P. Liang, "Semantic parsing on freebase from question-answer pairs," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1533–1544.

[11] M. Bhandari, P. N. Gour, A. Ashfaq, P. Liu, and G. Neubig, "Re-evaluating evaluation in text summarization," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 9347–9359. DOI: 10.18653/v1/2020.emnlp-main.751. [Online]. Available: https://aclanthology.org/2020.emnlp-main.751.

[12] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 2008, pp. 1247–1250.

[13] M. Braverman, X. Chen, S. M. Kakade, K. Narasimhan, C. Zhang, and Y. Zhang, "Calibration, entropy rates, and memory in language models," *arXiv preprint arXiv:1906.05664*, 2019.

[14] T. Brown *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[15] H. Bunt *et al.*, "The ISO standard for dialogue act annotation, second edition," English, in *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, May 2020, pp. 549–558, ISBN: 979-10-95546-34-4. [Online]. Available: https://aclanthology.org/2020.lrec-1.69.

[16] S. Cao and L. Wang, "CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6633–6649. DOI: 10.18653/v1/2021.emnlp-main.532. [Online]. Available: https://aclanthology.org/2021.emnlp-main.532.

[17] Z. Cao, F. Wei, W. Li, and S. Li, "Faithful to the original: Fact aware neural abstractive summarization," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11912.

[18] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. Conf. on Neural Information Processing Systems (NeurIPS)*, 2020.

[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[20] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[21] Z. Chen and T. Qian, "Transfer capsule network for aspect level sentiment classification," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 547–556.

[22] E. Choi *et al.*, "QuAC: Question answering in context," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 2174–2184. DOI: 10.18653/v1/D18-1241. [Online]. Available: https://aclanthology.org/D18-1241.

[23] K. W. Church and W. A. Gale, "Poisson mixtures," *Natural Language Engineering*, vol. 1, no. 2, pp. 163–190, 1995.

[24] C. D. Corley and R. Mihalcea, "Measuring the semantic similarity of texts," in *Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment*, 2005, pp. 13–18.

[25] T. De Jong, "Cognitive load theory, educational research, and instructional design: Some food for thought," *Instructional science*, vol. 38, no. 2, pp. 105–134, 2010.

[26] D. DeStefano and J.-A. LeFevre, "Cognitive load in hypertext reading: A review," *Computers in human behavior*, vol. 23, no. 3, pp. 1616–1641, 2007.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. [Online]. Available: https://www.aclweb.org/anthology/N19-1423.

[29] B. Dhingra, M. Faruqui, A. Parikh, M.-W. Chang, D. Das, and W. Cohen, "Handling divergent reference texts when evaluating table-to-text generation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4884–4895.

[30] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, "Wizard of wikipedia: Knowledge-powered conversational agents," in *International Conference on Learning Representations*, 2018.

[31] E. Dinan *et al.*, "The second conversational intelligence challenge (convai2)," *arXiv preprint arXiv:1902.00098*, 2019.

[32] E. Durmus, H. He, and M. Diab, "FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 5055–5070. DOI: 10.18653/v1/2020.acl-main.454. [Online]. Available: https://aclanthology.org/2020.acl-main.454.

[33] E. Durmus, F. Ladhak, and T. Hashimoto, "Spurious correlations in reference-free evaluation of text generation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1443–1454. [Online]. Available: https://aclanthology.org/2022.acl-long.102.

[34] O. Dušek, J. Novikova, and V. Rieser, "Findings of the e2e nlg challenge," in *Proceedings of the 11th International Conference on Natural Language Generation*, 2018, pp. 322–328.

[35] O. Dušek, J. Novikova, and V. Rieser, "Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge," *Computer Speech & Language*, vol. 59, pp. 123–156, 2020.

[36] N. Dziri, E. Kamalloo, K. Mathewson, and O. Zaiane, "Evaluating coherence in dialogue systems using entailment," in *Proceedings of the 2019 Workshop on Widening NLP*, Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 146–148. [Online]. Available: https://aclanthology.org/W19-3646.

[37] N. Dziri, E. Kamalloo, K. Mathewson, and O. Zaiane, "Evaluating coherence in dialogue systems using entailment," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 3806–3812. DOI: 10.18653/v1/N19-1381. [Online]. Available: https://www.aclweb.org/anthology/N19-1381.

[38] N. Dziri, E. Kamalloo, K. Mathewson, and O. R. Zaiane, "Augmenting neural response generation with context-aware topical attention," in *Proceedings of the First Workshop on NLP for Conversational AI*, 2019, pp. 18–31.

[39] N. Dziri, A. Madotto, O. Zaıane, and A. J. Bose, "Neural path hunter: Reducing hallucination in dialogue systems via path grounding," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2197–2214. DOI: 10.18653/v1/2021.emnlp-main.168. [Online]. Available: https://aclanthology.org/2021.emnlp-main.168.

[40] N. Dziri, S. Milton, M. Yu, O. Zaiane, and S. Reddy, "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*, Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 5271–5285. DOI: 10.18653/v1/2022.naacl-main.387. [Online]. Available: https://aclanthology.org/2022.naacl-main.387.

[41] N. Dziri, S. Milton, M. Yu, O. Zaıane, and S. Reddy, "On the origin of hallucinations in conversational models: Is it the datasets or the models?" *CoRR (arXiv preprint)*, vol. abs/2204.07931, 2022. DOI: 10.48550/arXiv.2204.07931.

arXiv: 2204.07931. [Online]. Available: https://doi.org/10.48550/arXiv.2204.07931.

[42] N. Dziri, H. Rashkin, T. Linzen, and D. Reitter, "Evaluating attribution in dialogue systems: The begin benchmark," *arXiv preprint arXiv:2105.00071*, 2021.

[43] N. Dziri *et al.*, "Faithdial: A faithful benchmark for information-seeking dialogue," *CoRR (arXiv preprint)*, vol. abs/2204.10757, 2022. DOI: 10.48550/ARXIV.2204.10757. [Online]. Available: https://arxiv.org/abs/2204.10757.

[44] P. Elias, "Predictive coding–i," *IRE Transactions on Information Theory*, vol. 1, no. 1, pp. 16–24, 1955.

[45] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, "SummEval: Re-evaluating summarization evaluation," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 391–409, 2021. DOI: 10.1162/tacl_a_00373. [Online]. Available: https://aclanthology.org/2021.tacl-1.24.

[46] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: improving visual-semantic embeddings with hard negatives," in *Proc. British Machine Vision Conference (BMVC)*, 2018, p. 12.

[47] T. Falke, L. F. R. Ribeiro, P. A. Utama, I. Dagan, and I. Gurevych, "Ranking generated summaries by correctness: An interesting but challenging application for natural language inference," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2214–2220. DOI: 10.18653/v1/P19-1213. [Online]. Available: https://aclanthology.org/P19-1213.

[48] A. Fan, M. Lewis, and Y. Dauphin, "Hierarchical neural story generation," *arXiv preprint arXiv:1805.04833*, 2018.

[49] S. Feng, E. Wallace, A. Grissom II, M. Iyyer, P. Rodriguez, and J. Boyd-Graber, "Pathologies of neural models make interpretations difficult," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3719–3728.

[50] D. Flickinger, E. M. Bender, and S. Oepen, *ERG semantic documentation*, Accessed on 2020-08-25, 2014. [Online]. Available: http://www.delph-in.net/esd.

[51] M. Freitag, G. Foster, D. Grangier, V. Ratnakar, Q. Tan, and W. Macherey, "Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1460–1474, Dec. 2021, ISSN: 2307-387X. DOI: 10.1162/tacl_a_00437. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00437/1979261/tacl_a_00437.pdf. [Online]. Available: https://doi.org/10.1162/tacl%5C_a%5C_00437.

[52] S. Gabriel, A. Celikyilmaz, R. Jha, Y. Choi, and J. Gao, "GO FIGURE: A meta evaluation of factuality in summarization," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online: Association for Computational Linguistics, Aug. 2021, pp. 478–487. DOI: 10.18653/v1/2021.findings-acl.42. [Online]. Available: https://aclanthology.org/2021.findings-acl.42.

[53] S. Gehrmann, E. Clark, and T. Sellam, "Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text," *CoRR (arXiv preprint)*, vol. abs/2202.06935, 2022. arXiv: 2202.06935. [Online]. Available: https://arxiv.org/abs/2202.06935.

[54] S. Gehrmann *et al.*, "The GEM benchmark: Natural language generation, its evaluation and metrics," in *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 96–120. DOI: 10.18653/v1/2021.gem-1.10. [Online]. Available: https://aclanthology.org/2021.gem-1.10.

[55] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.

[56] N. D. Goodman and M. C. Frank, "Pragmatic language interpretation as probabilistic inference," *Trends in cognitive sciences*, vol. 20, no. 11, pp. 818–829, 2016.

[57] K. Gopalakrishnan *et al.*, "Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations," in *Proc. Interspeech 2019*, 2019, pp. 1891–1895. DOI: 10.21437/Interspeech.2019-3079. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-3079.

[58] T. Goyal and G. Durrett, "Annotating and modeling fine-grained factuality in summarization," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, Jun. 2021, pp. 1449–1462. DOI: 10.18653/v1/2021.naacl-main.114. [Online]. Available: https://aclanthology.org/2021.naacl-main.114.

[59] P. Grice, *Studies in the Way of Words*. Harvard University Press, 1989.

[60] J. Grill *et al.*, "Bootstrap your own latent - A new approach to self-supervised learning," in *Proc. Conf. on Neural Information Processing Systems (NeurIPS)*, 2020.

[61] M. Grusky, M. Naaman, and Y. Artzi, "Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 708–719.

[62] P. Gupta, C.-S. Wu, W. Liu, and C. Xiong, "DialFact: A benchmark for fact-checking in dialogue," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3785–3801. DOI: 10.18653/v1/2022.acl-long.263. [Online]. Available: https://aclanthology.org/2022.acl-long.263.

[63] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 297–304.

[64] M. U. Gutmann and A. Hyvärinen, "Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics," *Journal of Machine Learning Research*, vol. 13, no. Feb, pp. 307–361, 2012.

[65] Y. Hao *et al.*, "An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 221–231.

[66] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced bert with disentangled attention," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=XPZIaotutsD.

[67] R. D. Hjelm *et al.*, "Learning deep representations by mutual information estimation and maximization," *arXiv preprint arXiv:1808.06670*, 2018.

[68] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," in *International Conference on Learning Representations*, 2019.

[69] M. Honnibal and I. Montani, "Spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing," *To appear*, vol. 7, no. 1, 2017.

[70] O. Honovich, L. Choshen, R. Aharoni, E. Neeman, I. Szpektor, and O. Abend, "Q^2: 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*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 7856–7870. DOI: 10.18653/v1/2021.emnlp-main.619. [Online]. Available: https://aclanthology.org/2021.emnlp-main.619.

[71] N. Houlsby *et al.*, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning*, 2019, pp. 2790–2799.

[72] A. Jabri, A. Owens, and A. A. Efros, "Space-time correspondence as a contrastive random walk," in *Proc. Conf. on Neural Information Processing Systems (NeurIPS)*, 2020.

[73] Z. Ji *et al.*, "Survey of hallucination in natural language generation," *arXiv preprint arXiv:2202.03629*, 2022.

[74] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "Spanbert: Improving pre-training by representing and predicting spans," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 64–77, 2020.

[75] M. Jovanović, M. Baez, and F. Casati, "Chatbots as conversational healthcare services," *IEEE Internet Computing*, vol. 25, no. 3, pp. 44–51, 2021. DOI: 10.1109/MIC.2020.3037151.

[76] D. Jurafsky and J. H. Martin, "Speech and language processing (draft)," *Chapter A: Hidden Markov Models (Draft of September 11, 2018). Retrieved March*, vol. 19, p. 2019, 2018.

[77] D. Kang and T. B. Hashimoto, "Improved natural language generation via loss truncation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 718–731. DOI: 10.18653/v1/2020.acl-main.66. [Online]. Available: https://aclanthology.org/2020.acl-main.66.

[78] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, "Ctrl: A conditional transformer language model for controllable generation," *arXiv preprint arXiv:1909.05858*, 2019.

[79] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.

[80] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980.

[81] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980.

[82] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[83] G. Klein, F. Hernandez, V. Nguyen, and J. Senellart, "The opennmt neural machine translation toolkit: 2020 edition," in *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, 2020, pp. 102–109.

[84] E. Kochmar, D. D. Vu, R. Belfer, V. Gupta, I. V. Serban, and J. Pineau, "Automated data-driven generation of personalized pedagogical interventions in intelligent tutoring systems," *International Journal of Artificial Intelligence in Education*, pp. 1–27, 2021. DOI: 10.1007/s40593-021-00267-x.

[85] P. Koehn and R. Knowles, "Six challenges for neural machine translation," in *Proceedings of the First Workshop on Neural Machine Translation*, 2017, pp. 28–39.

[86] L. Kong, C. d. M. d'Autume, W. Ling, L. Yu, Z. Dai, and D. Yogatama, "A mutual information maximization perspective of language representation learning," 2020.

[87] W. Kryscinski, B. McCann, C. Xiong, and R. Socher, "Evaluating the factual consistency of abstractive text summarization," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 9332–9346. DOI: 10.18653/v1/2020.emnlp-main.750. [Online]. Available: https://aclanthology.org/2020.emnlp-main.750.

[88] F. Ladhak, E. Durmus, H. He, C. Cardie, and K. McKeown, "Faithful or extractive? on mitigating the faithfulness-abstractiveness trade-off in abstractive summarization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1410–1421. DOI: 10.18653/v1/2022.acl-long.100. [Online]. Available: https://aclanthology.org/2022.acl-long.100.

[89] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.

[90] L. Laranjo *et al.*, "Conversational agents in healthcare: A systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 9, pp. 1248–1258, 2018. DOI: 10.1093/jamia/ocy072.

[91] K. Lee, O. Firat, A. Agarwal, C. Fannjiang, and D. Sussillo, "Hallucinations in neural machine translation," in *International Conference on Learning Representations*, 2019.

[92] M. Lewis *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. [Online]. Available: https://aclanthology.org/2020.acl-main.703.

[93] M. Lewis *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. [Online]. Available: https://aclanthology.org/2020.acl-main.703.

[94] C. Li, B. Bi, M. Yan, W. Wang, and S. Huang, "Addressing semantic drift in generative question answering with auxiliary extraction," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2021, pp. 942–947.

[95] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 110–119. DOI: 10.18653/v1/N16-1014. [Online]. Available: https://aclanthology.org/N16-1014.

[96] J. Li, M. Galley, C. Brockett, G. P. Spithourakis, J. Gao, and B. Dolan, "A persona-based neural conversation model," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2016, pp. 994–1003.

[97] J. Li, W. Monroe, and D. Jurafsky, "A simple, fast diverse decoding algorithm for neural generation," *arXiv preprint arXiv:1611.08562*, 2016.

[98] J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao, "Deep reinforcement learning for dialogue generation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1192–1202.

[99] M. Li *et al.*, "Don't say that! making inconsistent dialogue unlikely with unlikelihood training," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 4715–4728. DOI: 10.18653/v1/2020.acl-main.428. [Online]. Available: https://aclanthology.org/2020.acl-main.428.

[100] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013.

[101] Z. Lin, A. Madotto, Y. Bang, and P. Fung, "The adapter-bot: All-in-one controllable conversational model," *arXiv preprint arXiv:2008.12579*, 2020.

[102] Y. Liu *et al.*, "RoBERTa: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[103] K.-M. Lux, M. Sappelli, and M. Larson, "Truth or error? towards systematic analysis of factual errors in abstractive summaries," in *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, 2020, pp. 1–10.

[104] A. Madotto *et al.*, "Learning knowledge bases with parameters for task-oriented dialogue systems," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020, pp. 2372–2394.

[105] N. Mathur, T. Baldwin, and T. Cohn, "Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 4984–4997. DOI: 10.18653/v1/2020.acl-main.448. [Online]. Available: https://aclanthology.org/2020.acl-main.448.

[106] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 1906–1919.

[107] B. Mazoure, R. Tachet des Combes, T. Doan, P. Bachman, and R. D. Hjelm, "Deep reinforcement and infomax learning," 2020.

[108] D. McAllester, "Information theoretic co-training," 2018.

[109] D. McAllester and K. Stratos, "Formal limitations on the measurement of mutual information," *arXiv preprint arXiv:1811.04251*, 2018.

[110] T. McCoy, E. Pavlick, and T. Linzen, "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3428–3448. DOI: 10.18653/v1/P19-1334. [Online]. Available: https://aclanthology.org/P19-1334.

[111] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[112] G. A. Miller, "WordNet: A lexical database for English," in *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994. [Online]. Available: https://aclanthology.org/H94-1111.

[113] J. Min, R. T. McCoy, D. Das, E. Pitler, and T. Linzen, "Syntactic data augmentation increases robustness to inference heuristics," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 2339–2352. DOI: 10.18653/v1/2020.acl-main.212. [Online]. Available: https://aclanthology.org/2020.acl-main.212.

[114] S. Moon, P. Shah, A. Kumar, and R. Subba, "OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 845–854. DOI: 10.18653/v1/P19-1081. [Online]. Available: https://aclanthology.org/P19-1081.

[115] M. Nadeem, A. Bethke, and S. Reddy, "StereoSet: Measuring stereotypical bias in pretrained language models," 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)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 5356–5371. DOI: 10.18653/v1/2021.acl-long.416. [Online]. Available: https://aclanthology.org/2021.acl-long.416.

[116] R. Nakamura, K. Sudoh, K. Yoshino, and S. Nakamura, "Another diversity-promoting objective function for neural dialogue generation," in *AAAI Workshop on Reasoning and Learning for Human-Machine Dialogues (DEEP-DIAL)*, 2018.

[117] F. Nan *et al.*, "Improving factual consistency of abstractive summarization via question answering," 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)*, Online: Association for Computational Linguistics, 2021, pp. 6881–6894. DOI: 10.18653/v1/2021.acl-long.536. [Online]. Available: https://aclanthology.org/2021.acl-long.536.

[118] A. Newell, J. C. Shaw, and H. A. Simon, "Report on a general problem solving program," in *IFIP congress*, Pittsburgh, PA, vol. 256, 1959, p. 64.

[119] A. Newell and H. Simon, "The logic theory machine–a complex information processing system," *IRE Transactions on information theory*, vol. 2, no. 3, pp. 61–79, 1956.

[120] Y. Nie, M. Williamson, M. Bansal, D. Kiela, and J. Weston, "I like fish, especially dolphins: Addressing contradictions in dialogue modeling," 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)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 1699–1713. DOI: 10.18653/v1/2021.acl-long.134. [Online]. Available: https://aclanthology.org/2021.acl-long.134.

[121] Z.-Y. Niu, H. Wu, H. Wang, *et al.*, "Knowledge aware conversation generation with explainable reasoning over augmented graphs," 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)*, 2019, pp. 1782–1792.

[122] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. European Conf. on Computer Vision*, Springer, 2016, pp. 69–84.

[123] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[124] A. Pagnoni, V. Balachandran, and Y. Tsvetkov, "Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, Jun. 2021, pp. 4812–4829. DOI: 10.18653/v1/2021.naacl-main.383. [Online]. Available: https://aclanthology.org/2021.naacl-main.383.

[125] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, 2002, pp. 311–318.

[126] A. Parikh *et al.*, "ToTTo: A controlled table-to-text generation dataset," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 1173–1186. DOI: 10.18653/v1/2020.emnlp-main.89. [Online]. Available: https://aclanthology.org/2020.emnlp-main.89.

[127] F. Petroni *et al.*, "Language models as knowledge bases?" In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 2463–2473.

[128] B. Poole, S. Ozair, A. v. d. Oord, A. A. Alemi, and G. Tucker, "On variational bounds of mutual information," 2019.

[129] S. Prabhumoye, K. Hashimoto, Y. Zhou, A. W. Black, and R. Salakhutdinov, "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*, Online: Association for Computational Linguistics, Jun. 2021, pp. 4274–4287. DOI: 10.18653/v1/2021.naacl-main.338. [Online]. Available: https://aclanthology.org/2021.naacl-main.338.

[130] A. Pu, H. W. Chung, A. Parikh, S. Gehrmann, and T. Sellam, "Learning compact metrics for MT," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 751–762. DOI: 10.18653/v1/2021.emnlp-main.58. [Online]. Available: https://aclanthology.org/2021.emnlp-main.58.

[131] J. T. Quah and Y. Chua, "Chatbot assisted marketing in financial service industry," in *International Conference on Services Computing*, Springer, 2019, pp. 107–114.

[132] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[133] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html.

[134] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016. [Online]. Available: http://arxiv.org/abs/1511.06732.

[135] H. Rashkin, D. Reitter, G. S. Tomar, and D. Das, "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)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 704–718. DOI: 10.18653/v1/2021.acl-long.58. [Online]. Available: https://aclanthology.org/2021.acl-long.58.

[136] H. Rashkin *et al.*, "Measuring attribution in natural language generation models," *arXiv preprint arXiv:2112.12870*, 2021.

[137] V. Raunak, A. Menezes, and M. Junczys-Dowmunt, "The curious case of hallucinations in neural machine translation," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, Jun. 2021, pp. 1172–1183. DOI: 10.18653/v1/2021.naacl-main.92. [Online]. Available: https://aclanthology.org/2021.naacl-main.92.

[138] S. Reddy, D. Chen, and C. D. Manning, "CoQA: A conversational question answering challenge," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 249–266, 2019. DOI: 10.1162/tacl_a_00266. [Online]. Available: https://aclanthology.org/Q19-1016.

[139] M. T. Review, "The pandemic is emptying call centers. ai chatbots are swooping in.," *https://www.technologyreview.com/2020/05/14/1001716/ai-chatbots-take-call-center-jobs-during-coronavirus-pandemic/*, 2020.

[140] B. Rhodes, K. Xu, and M. U. Gutmann, "Telescoping density-ratio estimation," *arXiv preprint arXiv:2006.12204*, 2020.

[141] A. Roberts, C. Raffel, and N. Shazeer, "How much knowledge can you pack into the parameters of a language model?" In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 5418–5426.

[142] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko, "Object hallucination in image captioning," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 4035–4045.

[143] S. Roller *et al.*, "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*, Online: Association for Computational Linguistics, Apr. 2021, pp. 300–325. DOI: 10.18653/v1/2021.eacl-main.24. [Online]. Available: https://aclanthology.org/2021.eacl-main.24.

[144] S. Santhanam *et al.*, "Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation," *CoRR*, vol. abs/2110.05456, 2021. arXiv: 2110.05456. [Online]. Available: https://arxiv.org/abs/2110.05456.

[145] A. See, S. Roller, D. Kiela, and J. Weston, "What makes a good conversation? how controllable attributes affect human judgments," *arXiv preprint arXiv:1902.08654*, 2019.

[146] A. See, S. Roller, D. Kiela, and J. Weston, "What makes a good conversation? how controllable attributes affect human judgments," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1702–1723. DOI: 10.18653/v1/N19-1170. [Online]. Available: https://aclanthology.org/N19-1170.

[147] T. Sellam, D. Das, and A. Parikh, "BLEURT: Learning robust metrics for text generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 7881–7892. DOI: 10.18653/v1/2020.acl-main.704. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.704.

[148] I. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, 2016.

[149] K. B. Sheehan, "Crowdsourcing research: Data collection with amazon's mechanical turk," *Communication Monographs*, vol. 85, no. 1, pp. 140–156, 2018.

[150] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval augmentation reduces hallucination in conversation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3784–3803. DOI: 10.18653/v1/2021.findings-emnlp.320. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.320.

[151] A. Singhal, "Introducing the knowledge graph: Things, not strings," *Official google blog*, vol. 5, 2012.

[152] R. Sipos, P. Shivaswamy, and T. Joachims, "Large-margin learning of submodular summarization models," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, pp. 224–233.

[153] J. Song and S. Ermon, "Understanding the limitations of variational mutual information estimators," 2019.

[154] J. Song and S. Ermon, "Multi-label contrastive predictive coding," 2020.

[155] A. Sordoni, N. Dziri, H. Schulz, G. Gordon, P. Bachman, and R. T. Des Combes, "Decomposed mutual information estimation for contrastive representation learning," in *International Conference on Machine Learning*, PMLR, 2021, pp. 9859–9869.

[156] A. Sordoni *et al.*, "A neural network approach to context-sensitive generation of conversational responses," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 196–205.

[157] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[158] W. B. Stiles, *Describing talk: A taxonomy of verbal response modes*. Sage Publications, 1992.

[159] K. Stratos, "Mutual information maximization for simple and accurate part-of-speech induction," *arXiv preprint arXiv:1804.07849*, 2018.

[160] S. F. Suhel, V. K. Shukla, S. Vyas, and V. P. Mishra, "Conversation to automation in banking through chatbot using artificial machine intelligence language," in *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, IEEE, 2020, pp. 611–618.

[161] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *neural information processing systems*, pp. 3104–3112, 2014.

[162] X. Tang *et al.*, "CONFIT: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning," *arXiv preprint arXiv:2112.08713*, 2021.

[163] W. L. Taylor, ""cloze procedure": A new tool for measuring readability," *Journalism quarterly*, vol. 30, no. 4, pp. 415–433, 1953.

[164] R. Thoppilan *et al.*, "Lamda: Language models for dialog applications," *arXiv preprint arXiv:2201.08239*, 2022.

[165] R. Tian, S. Narayan, T. Sellam, and A. P. Parikh, "Sticking to the facts: Confident decoding for faithful data-to-text generation," *CoRR (arXiv preprint)*, vol. abs/1910.08684, 2019. arXiv: 1910.08684. [Online]. Available: http://arxiv.org/abs/1910.08684.

[166] R. Tian, S. Narayan, T. Sellam, and A. P. Parikh, "Sticking to the facts: Confident decoding for faithful data-to-text generation," *CoRR*, vol. abs/1910.08684, 2019. arXiv: 1910.08684. [Online]. Available: http://arxiv.org/abs/1910.08684.

[167] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning," *arXiv preprint arXiv:2005.10243*, 2020.

[168] Y.-L. Tuan, Y.-N. Chen, and H.-y. Lee, "Dykgchat: Benchmarking dialogue generation grounding on dynamic knowledge graphs," 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)*, 2019, pp. 1855–1865.

[169] S. Vashishth, S. Sanyal, V. Nitin, and P. Talukdar, "Composition-based multi-relational graph convolutional networks," *arXiv preprint arXiv:1911.03082*, 2019.

[170] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[171] P. Velickovic, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, "Deep graph infomax," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2019.

[172] A. Wang, K. Cho, and M. Lewis, "Asking and answering questions to evaluate the factual consistency of summaries," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 5008–5020.

[173] C. Wang and R. Sennrich, "On exposure bias, hallucination and domain shift in neural machine translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 3544–3552. DOI: 10.18653/v1/2020.acl-main.326. [Online]. Available: https://aclanthology.org/2020.acl-main.326.

[174] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 28, 2014.

[175] S. Welleck, I. Kulikov, S. Roller, E. Dinan, K. Cho, and J. Weston, "Neural text generation with unlikelihood training," *arXiv preprint arXiv:1908.04319*, 2019.

[176] S. Welleck, J. Weston, A. Szlam, and K. Cho, "Dialogue natural language inference," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3731–3741. DOI: 10.18653/v1/P19-1363. [Online]. Available: https://www.aclweb.org/anthology/P19-1363.

[177] S. Welleck, J. Weston, A. Szlam, and K. Cho, "Dialogue natural language inference," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3731–3741. DOI: 10.18653/v1/P19-1363. [Online]. Available: https://aclanthology.org/P19-1363.

[178] S. Welleck, J. Weston, A. Szlam, and K. Cho, "Dialogue natural language inference," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3731–3741.

[179] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 1112–1122. DOI: 10.18653/v1/N18-1101. [Online]. Available: https://aclanthology.org/N18-1101.

[180] R. J. Williams and D. Zipser, "Experimental analysis of the real-time recurrent learning algorithm," *Connection science*, vol. 1, no. 1, pp. 87–111, 1989.

[181] G. N. Wire, "Chatbot market growth is projected to reach usd 3.62 billion by 2030, growing at a cagr of 23.9%: Straits research," *Globenewswire*, 2022.

[182] S. Wiseman, S. Shieber, and A. Rush, "Challenges in data-to-document generation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2253–2263. DOI: 10.18653/v1/D17-1239. [Online]. Available: https://aclanthology.org/D17-1239.

[183] T. Wolf, V. Sanh, J. Chaumond, and C. Delangue, "Transfertransfo: A transfer learning approach for neural network based conversational agents," *CoRR (arXiv preprint)*, vol. abs/1901.08149, 2019. arXiv: 1901.08149. [Online]. Available: http://arxiv.org/abs/1901.08149.

[184] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.6.

[185] H. Xu, S. Moon, H. Liu, B. Liu, P. Shah, and S. Y. Philip, "User memory reasoning for conversational recommendation," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 5288–5308.

[186] S. Yang and C. Evans, "Opportunities and challenges in using ai chatbots in higher education," in *Proceedings of the 2019 3rd International Conference on Education and E-Learning*, ser. ICEEL 2019, Barcelona, Spain: Association for Computing Machinery, 2019, pp. 79–83, ISBN: 9781450372251. DOI: 10.1145/3371647.3371659. [Online]. Available: https://doi.org/10.1145/3371647.3371659.

[187] X. Yao and B. Van Durme, "Information extraction over structured data: Question answering with freebase," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 956–966.

[188] Y.-T. Yeh, M. Eskenazi, and S. Mehri, "A comprehensive assessment of dialog evaluation metrics," in *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, Online: Association for Computational Linguistics, Nov. 2021, pp. 15–33. DOI: 10.18653/v1/2021.eancs-1.3. [Online]. Available: https://aclanthology.org/2021.eancs-1.3.

[189] T. Young, E. Cambria, I. Chaturvedi, H. Zhou, S. Biswas, and M. Huang, "Augmenting end-to-end dialogue systems with commonsense knowledge," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[190] M. Yu and M. Dredze, "Improving lexical embeddings with semantic knowledge," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2014, pp. 545–550.

[191] W. Yuan, G. Neubig, and P. Liu, "Bartscore: Evaluating generated text as text generation," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., 2021, pp. 27 263–27 277. [Online]. Available: https://proceedings.neurips.cc/paper/2021/file/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Paper.pdf.

[192] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "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)*, Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2204–2213. DOI: 10.18653/v1/P18-1205. [Online]. Available: https://aclanthology.org/P18-1205.

[193] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "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)*, Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2204–2213. DOI: 10.18653/v1/P18-1205. [Online]. Available: https://www.aclweb.org/anthology/P18-1205.

[194] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2019.

[195] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr.

[196] Y. Zhang *et al.*, "Generating informative and diverse conversational responses via adversarial information maximization," 2018, pp. 1810–1820.

[197] Y. Zhang *et al.*, "DIALOGPT : Large-scale generative pre-training for conversational response generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Online: Association for Computational Linguistics, Jul. 2020, pp. 270–278. DOI: 10.18653/v1/2020.acl-demos.30. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-demos.30.

[198] Y. Zhang *et al.*, "DIALOGPT : Large-scale generative pre-training for conversational response generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Online: Association for Computational Linguistics, Jul. 2020, pp. 270–278. DOI: 10.18653/v1/2020.acl-demos.30. [Online]. Available: https://aclanthology.org/2020.acl-demos.30.

[199] Y. Zhang *et al.*, "Dialogpt: Large-scale generative pre-training for conversational response generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020, pp. 270–278.

[200] H. Zhou, T. Young, M. Huang, H. Zhao, J. Xu, and X. Zhu, "Commonsense knowledge aware conversation generation with graph attention.," in *IJCAI*, 2018, pp. 4623–4629.

[201] K. Zhou, S. Prabhumoye, and A. W. Black, "A dataset for document grounded conversations," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 708–713. DOI: 10.18653/v1/D18-1076. [Online]. Available: https://aclanthology.org/D18-1076.

[202] W. Zhu, K. Mo, Y. Zhang, Z. Zhu, X. Peng, and Q. Yang, "Flexible end-to-end dialogue system for knowledge grounded conversation," *arXiv*, arXiv–1709, 2017.