Attention Biasing and Context Augmentation for Zero-Shot Control of Encoder-Decoder Transformers for Natural Language Generation

Devamanyu Hazarika [2*], **Mahdi Namazifar** [1], **Dilek Hakkani-Tür** [1]

[1] Amazon Alexa AI
[2] National University of Singapore
hazarika@comp.nus.edu.sg, {mahdinam,hakkanit}@amazon.com

Abstract

Controlling neural network-based models for natural language generation (NLG) to realize desirable attributes in the generated outputs has broad applications in numerous areas such as machine translation, document summarization, and dialog systems. Approaches that enable such control in a *zero-shot* manner would be of great importance as, among other reasons, they remove the need for additional annotated data and training. In this work, we propose novel approaches for controlling encoder-decoder transformer-based NLG models in zero shot. While zero-shot control has previously been observed in massive models (e.g., GPT3), our method enables such control for smaller models. This is done by applying two control knobs, attention biasing and context augmentation, to these models directly during decoding and without additional training or auxiliary models. These knobs control the generation process by directly manipulating trained NLG models (e.g., biasing cross-attention layers). We show that not only are these NLG models robust to such manipulations, but also their behavior could be controlled without an impact on their generation performance.

1 Introduction

Natural language generation (NLG) aims at producing fluent and coherent sentences and phrases in different problem settings such as dialog systems (Huang, Zhu, and Gao 2020), machine translation (Yang, Wang, and Chu 2020), and text summarization (Syed, Gaol, and Matsuo 2021). Most recently, the majority of the research in NLG leverages transformers (Vaswani et al. 2017) and specifically transformer decoders to generate natural language (Radford et al. 2019; Brown et al. 2020; Lewis et al. 2020). Although these statistical approaches to NLG have proven to be highly effective, their stochastic nature and complex architectures make them difficult to control in order for them to reflect any set of desired attributes in the output. These attributes could range from persona, sentiment, empathy, dialog acts for dialog response generation (Niu and Bansal 2018; Zhang et al. 2018; See et al. 2019; Madotto et al. 2020) to story ending control for story generation (Peng et al. 2018) or formality and politeness control for drafting emails (Madaan et al. 2020).

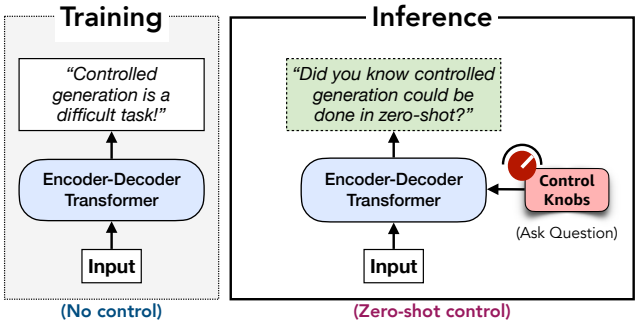

Figure 1: Zero-Shot Controlled Generation on an Encoder-Decoder Transformer at inference time using Control Knobs. Control knobs influence the generation process such that the output has the desired attributes (e.g., asking questions).

In general, being able to control an NLG model in a *zero-shot* fashion would be highly instrumental since such zero-shot control would *not* require large amounts of attribute annotated data, nor would it require any control-specific fine-tuning of the NLG model or auxiliary attribute models to guide the generation (Dathathri et al. 2020; Yang and Klein 2021). However, zero-shot control of NLG models is a non-trivial task with multiple challenges. Previous zero-shot works only control lexical constraints, like controlling token diversity or sentence length using decoding heuristics (Keskar et al. 2019; Vijayakumar et al. 2016; Holtzman et al. 2020), and thus cannot be extended to control a broader range of attributes. Furthermore, encoder-decoder transformer NLG (EDT-NLG) models are often used for grounded (or directed) NLG tasks, like dialog response generation and summarization, requiring the models to be fluent in their generation and relevant to the conditioning input. As such, adding requirements for zero-shot control during generation increases the complexity of the task as the model has to achieve the control without compromising its fluency and relevance. Given this reduced degree of freedom, balancing the trade-off between the amount of control and quality of the generations is extremely difficult, particularly since the model is trained only for the latter and introduced with the former only during inference (testing) phase.

In this work, we explore the challenges mentioned above and introduce new zero-shot approaches for controlling

*This work was done when Devamanyu interned at Amazon Alexa AI.

EDT-NLG models. The high-level idea of our approach is to explicitly manipulate the transformers within the already trained EDT-NLG models to achieve the desired attributes at generation time (see Fig. 1). More specifically, we introduce two control knobs, *attention biasing* and *context augmentation*, that could be used to control the generation of EDT-NLG models in zero shot. The attention biasing knob modifies the attention paid to different parts of the context provided as input. This is done by directly manipulating the cross-attention distributions generated by the EDT-NLG model during inference. As context often comprises multiple components (examples in dialog systems include different speaker turns, knowledge snippets, knowledge graphs, images, etc.), we demonstrate that biasing attention towards specific components reveals predictable control on the model's generations. Crucially, we find this control in zero shot, directly at inference time.

On the other hand, the context augmentation knob works by introducing additional context on the encoder side. This enables the model to condition upon additional attributes that are not part of the original context, such as sentiment, style, topics, etc. Similar to the attention biasing, the model is not trained to condition on these novel contexts and instead introduced in a zero-shot manner directly during inference. As the word "*knob*" suggests, the design of our control knobs is very flexible and allows varying degrees of control.

In summary, our contributions are as follows:

- We empirically show that manually interfering with cross-attention distributions within already trained transformers does *not* derail the model's generative capabilities, and we can leverage this robustness to control the behavior of these models in zero shot.
- We propose two control knobs that can control EDT-NLG models during generation in a zero-shot manner, i.e., without control-specific training, using any attribute discriminator, or gradient-based optimization during inference.
- We demonstrate that by combining the knobs, we can achieve zero-shot control on even small transformer models like BART base (Lewis et al. 2020). Such zero-shot control was previously observed only in huge transformer models like GPT3 (Brown et al. 2020).
- We apply the proposed control knobs to the Knowledge-Grounded Neural Response Generation (K-NRG) task and find that these control knobs can manifest a wide variety of attributes in zero shot, that includes improving *informativeness*, *inquisitiveness* (asking engaging questions), *positive sentiment* of the responses, amongst others.

2 Related Works

Numerous works in the literature focus on controlling NLG models (Prabhumoye, Black, and Salakhutdinov 2020), which fall under two major categories. The first category focuses on using data annotated with the desired attributes to train the NLG model such that it can generate with the same attributes (Keskar et al. 2019; Wu et al. 2020; Smith et al. 2020; See et al. 2019; Rashkin et al. 2021). The drawback of this approach is that for every set of attributes, annotated datasets are required, which makes the approach difficult to scale. The second category of approaches achieves the desired control either using attribute discriminators (generative (Krause et al. 2020) or discriminative (Yang and Klein 2021)) or bag-of-words that are indicative of the attributes (Ghazvininejad et al. 2017; Baheti et al. 2018; See et al. 2019). However, these decoding strategies have been observed to be brittle, particularly for tasks like dialog response generation (See et al. 2019). Another set of approaches within this category, namely Plug-and-Play Language Models (PPLM) (Dathathri et al. 2020) leverage gradients of auxiliary models that can detect the desired attributes. In these methods, training auxiliary models still require annotated data that could be expensive to acquire. Moreover, PPLMs are computationally expensive as they employ gradient updates for each token during generation.

In contrast to the above categories, the goal of this work is to control NLG in zero shot. Along this goal, prompt-based approaches have been proposed that prime massive language models, like GPT-3 (Brown et al. 2020), with few-shot supervised examples of a specific task. Recently (Schick and Schütze 2020) enabled such behavior in smaller models, however not in the zero-shot setting. To the best of our knowledge, there is no work in the literature focused on controlling the output within a grounded NLG task (Wu et al. 2020) in a zero-shot setting.

3 Zero-Shot Control Knobs for NLG

Fig. 2 describes an EDT-NLG model π_θ with trained parameters θ and conditioned on an input (or context) x. The model first encodes this input and computes the encoded context (enc(x)). The decoder then generates the output y by sampling one token at a time in an auto-regressive manner, i.e., $\hat{y} \sim \pi_\theta(y|x)$, and grounds the output to the encoded context using a cross-attention mechanism. Now, generating with additional desired attributes, e.g., positive sentiment, could be interpreted as introducing an additional condition c to the sampling process. The control knobs introduced in this work (attention biasing and context augmentation) manually modify π_θ to $\tilde{\pi}_{\tilde{\theta}}$ during generation, such that samples from $\tilde{\pi}_{\tilde{\theta}}(y|x, c)$ on average manifest the desired attributes significantly more than samples from π_θ. Note that throughout this process, $\pi_\theta(y|x, c)$ is never trained.

3.1 Attention Biasing

Consider the cross-attention layer in the decoder of an EDT-NLG model. At generation time step t, the decoder attends to the encoded input in the following manner: the query vector is first multiplied by the key matrix, and the result goes through a Softmax operation that outputs a discrete probability distribution, that is referred to as *attention distribution*. Attention distribution is then used (through multiplication with the value matrix) to determine, in some sense, how much attention should be paid to each one of the attention context tokens (Daniluk et al. 2017). The idea of the attention biasing knob (ATTN. BIAS) is forcing an attention module to attend to some parts of its context more (or less) than it normally would. For example, in the task of dialog response generation, perhaps the input might in-

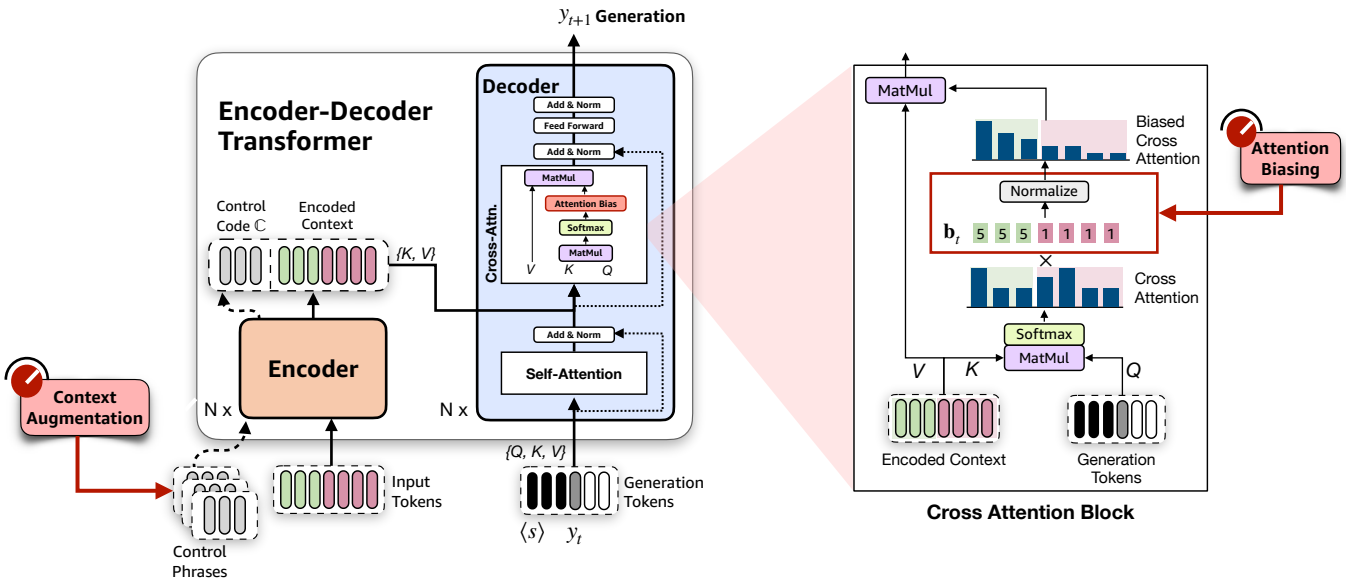

Figure 2: Control knobs for zero-shot controlled NLG: attention biasing and context augmentation knobs. Right part of the image depicts the attention biasing knob for cross-attention. Values 1 and 5 are example bias values. Best viewed in color.

clude a knowledge snippet in addition to conversation history and we would like the generated response to fully incorporate the knowledge snippet. This is done through the ATTN. BIAS knob by directly adjusting the attention distribution in cross-attention of EDT-NLG models.

The ATTN. BIAS knob works through element-wise multiplication of a *bias vector* with the attention distribution and then normalizing the results so that the outcome is still a probability distribution (referred to as biased attention distribution). As an example, in Fig. 2 (right part), the cross-attention context has two parts, and the attention process is biased by multiplying the attention to the first part by some value (for example, 5) and then normalizing the outcome to retrieve a probability distribution[1]. This particular example emphasizes the first part of the context over the second. More formally, given embedded attention context $C = \text{enc}(x)$, attention matrices $W_K, W_V, W_Q \in \mathbb{R}^{d \times d}$, and the embedding $\mathbf{e}_t \in \mathbb{R}^d$ for y_t, cross-attention output for y_t is:

$$\text{softmax}\left(\frac{(\mathbf{e}_t W_Q)(CW_K)^T}{\sqrt{d}}\right) CW_V$$

In this notation, biased cross-attention could be defined as:

$$\mathcal{N}\left(\mathbf{b}_t \odot \text{softmax}\left(\frac{(\mathbf{e}_t W_Q)(CW_K)^T}{\sqrt{d}}\right)\right) CW_V,$$

where function $\mathcal{N}$ normalizes a given positive vector to have the element-wise sum of 1, $\mathbf{b}_t$ is the bias vector at time step t, and $\odot$ represents element-wise vector multiplication.

Training with biased attention modules has been employed in tasks like machine translation, to achieve focused attention (Luong, Pham, and Manning 2015; Yang et al. 2018; You, Sun, and Iyyer 2020; Shaw, Uszkoreit, and Vaswani 2018). However, these works do not employ bias in a zero-shot setting. Zero-shot biases have been studied in the probing literature to understand the influence of attention on model's classifications (Serrano and Smith 2019), but to the best of our knowledge, zero-shot attention biasing for controlled generation is an unexplored avenue.

Note that in this work vector $\mathbf{b}_t$ is not a learned parameter, and it is set manually. For applications like dialog responses, higher-level planners or dialog managers (Hedayatnia et al. 2020; Rashkin et al. 2021) could be responsible for determining the value of $\mathbf{b}_t$. However, here we dedicate our focus on establishing the feasibility of controlled generation through attention biasing, and leave the question of how to determine the amount of bias for future works.

3.2 Context Augmentation

In the *context augmentation* knob (CTX. AUG.), we apply modifications to the input of the EDT-NLG model in order to push the model to manifest the desired attributes in the generations. We explain how the knob works through an example. Imagine that the desired attribute for the output of the model is *asking a question*, i.e., inquisitive generation (Fig. 1). For this, we would like to increase the likelihood of the model's output towards including a question. To this end, we first sample a set of question sentences (e.g., by choosing sentences that end with a question mark) from any text corpora. We call these sentences *control phrases*. We then feed each control phrase to the encoder of the EDT-NLG model to get an embedding for it.We then take the average of these embeddings across all control phrases, which we refer to as *control code* and we denote it as $\mathbb{C}$. The control code $\mathbb{C}$ is then concatenated ($\oplus$) to the encoded context: $\mathbb{C} \oplus \text{enc}(x)$, as shown in Fig. 2. $\mathbb{C} \oplus \text{enc}(x)$ is then used (attended to) by

[1]Our initial experiments also explored biasing the decoder self-attention in zero shot, but we found that it degenerates the output text. Deeper dive into biasing decoder self-attention in zero shot is out of the scope of this work.

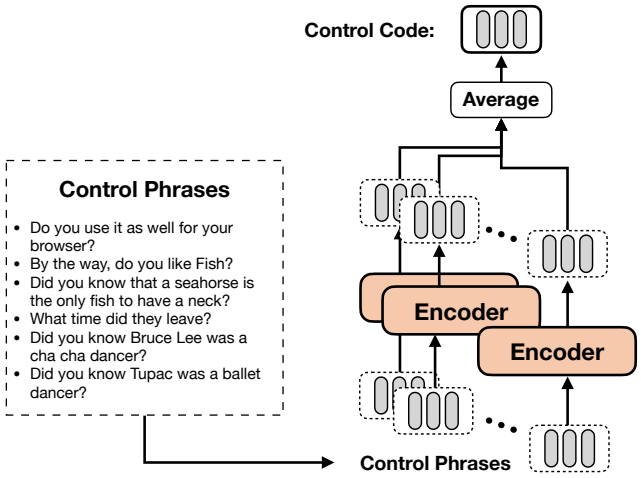

Figure 3: Example of creating control codes for questions.

the decoder. Note that without CTX. AUG. knob, the decoder uses only enc(x) at generation time. The overall process of generating control codes is illustrated in Fig. 3.

The control code, inspired from *prototypes* (Snell, Swersky, and Zemel 2017), is designed to capture the shared concepts among the control phrases. The role of averaging in creating control codes is to maintain the shared concepts within the control phrases and smoothing out other concepts, such as topic.

4 Experiments

The control knobs introduced in this paper could be generalized to any EDT-NLG model for any grounded NLG task. For example, the ATTN. BIAS knob is generic to any attention mechanism within or outside of a transformer-based architecture and could be utilized in other attention-based applications such as vision and multi-modal problems in zero shot. With that being said, to demonstrate the efficacy of the control knobs, we focus on a specific family of NLG tasks, namely knowledge-grounded open-domain Neural Response Generation (K-NRG).

4.1 Preliminaries

For the K-NRG task, we train an EDT-NRG model π_θ (without control-specific training). At every turn, the input dialog context x comprises of the previous dialog turns h, concatenated to the knowledge snippet k, i.e. $x = (k, h)$ (refer to Table 3 for an example of h and k). The decoder is prompted with $\langle s \rangle$ token and in an auto-regressive manner generates one token (y_{t+1}) at a time until a special end-of-sentence token is generated, i.e., $y_{t+1} \sim \pi_\theta(y|\mathbf{k}, \mathbf{h}, y_1, ..., y_t)$.

Data Setup. For our experiments, we use the Topical-Chat dataset (Gopalakrishnan et al. 2019) which includes dialogues between two Mechanical Turkers. Similar to related works in the literature (Hedayatnia et al. 2020; Rashkin et al. 2021), we assume that for each dialog context, the knowledge snippet is available, which is retrieved using TF-IDF similarity to the ground truth response. We evaluate the efficacy of the control knobs over the two *frequent* and *rare* test sets from the Topical-Chat dataset.

Model Setup. For the NRG model, we use BART-base as the pre-trained EDT model (Lewis et al. 2020). We choose the smaller model, as it is more difficult to achieve zero-shot control in smaller models (Schick and Schütze 2020). Additionally, smaller models are more economical with a much less carbon footprint

Input Setup. The input to the K-NRG problem comprises of a knowledge snippet k and the dialog history h which has the last five turns in the dialog, with respect to the response. We assign a fixed number of tokens for each sentence in the input, and infill with empty pads if required. In particular, for k we provide 32 tokens and 25 tokens for each turn in h.

The overall dialog context starts with the special token $\langle s \rangle$, followed by k. Next, we include the dialog history, whose turns use alternate start symbols: $\langle speaker1 \rangle, \langle speaker2 \rangle$. Overall, our input to the model (dialog context) is composed of 163 tokens (33 knowledge tokens plus 26 turn tokens for each of the 5 turns).

Training and Inference. We train the BART-base model with maximum likelihood-based training using ground-truth human responses – this process does not use any control knobs. We train for a maximum of 10 epochs with early stopping (patience = 1 on average perplexity of validation set). We train with a batch size of 5, gradient accumulation of 4, and learning rate of $6.25e-5$. For inference, we follow (Hedayatnia et al. 2020) and utilize nucleus sampling (Holtzman et al. 2020) with a top-p value of 0.9. Top-k is set to 0 and temperature is set to 0.7. The maximum length of the responses is set to 40 tokens.

Goals of the Experiments. The goals of our experiments are two-fold. First, we examine whether the proposed control knobs effectively control the generation as per the desired attributes. Second, we examine whether applying the knobs would cause negative impacts (trade-offs) on the generation output. Specifically, we examine the impact of the control knobs on *fluency* and *relevance* of the generated response. In the literature, fluency refers to the grammatical and syntactical correctness of generated responses, and relevance refers to appropriateness of a response given the history of the dialog (See et al. 2019; Shin et al. 2019; Rashkin et al. 2019). We repeat our experiments across five runs to account for variability in the token sampling procedure.

4.2 Attention Biasing Experiments

In this section, we study the effects of applying cross-attention biasing (§ 3.1) for zero-shot control of *informativeness* in generated responses for the K-NRG task. We apply the ATTN. BIAS knob to the cross-attention modules of an EDT-NRG model fine-tuned on Topical-Chat.

Given a dialog context $x = (k, h)$, the bias vector at generation time step t could be represented as $\mathbf{b}_t$ which is the

Knob	Bias Profile	Fluency			Relevance			Informativeness						
		$\textbf{PPL}_r$		Human Eval	$\textbf{BERTScore}_r$		Human Eval	$\textbf{BLEU}_k$		$\textbf{ROUGE}_k$		$\textbf{METEOR}_k$		Human Eval
		Freq	Rare	$[0, 1]$	Freq	Rare	$[1, 5]$	Freq	Rare	Freq	Rare	Freq	Rare	$[1, 5]$
Base Model		9.66	9.88	0.796	0.27	0.27	3.76	0.09	0.16	0.22	0.28	0.28	0.36	3.43
ATTN. BIAS	*Dialog*	10.15	10.39	-	0.24	0.24	-	**0.03**	**0.10**	**0.13**	**0.20**	**0.16**	**0.26**	-
	Knowledge	10.20	10.59	0.786	0.27	0.27	3.81	**0.14**	**0.26**	**0.28**	**0.38**	**0.36**	**0.49**	**3.84**
	Gradual-Knowledge	10.03	10.38	0.788	0.27	0.27	3.83	**0.13**	**0.22**	**0.26**	**0.34**	**0.34**	**0.45**	**3.80**

Table 1: Effect of ATTN. BIAS control knob on the informativeness of responses for the Topical-Chat *frequent* and *rare* test sets. Numbers in boldface represent statistically significant difference with respect to Base Model as per both pairwise Tukey's HSD test and two-tailed unpaired t-test (both with $p < 0.001$ over five independent runs). We skip the human evaluations for Dialog profile as it does not aid in improving the desired skill of informativeness in responses.

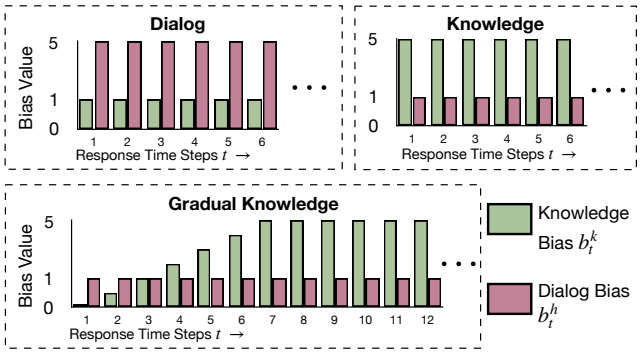

Figure 4: Cross-attention biasing profiles: Dialog, Knowledge, and Gradual Knowledge. Best viewed in color.

concatenation of two bias vectors $\mathbf{b}_t^k$ and $\mathbf{b}_t^h$ (see Fig. 2)[2]:

$$\mathbf{b}_t = [\mathbf{b}_t^k; \mathbf{b}_t^h] \quad s.t. \quad \mathbf{b}_t^k = \left[\left(b_t^k\right)_{\times |k|}\right],\ \mathbf{b}_t^h = \left[\left(b_t^h\right)_{\times |h|}\right],$$

where, $|k|$ and $|h|$ represent the total number of tokens in knowledge and dialog history, respectively. For example, if k has 3 tokens and h has 4 tokens, and at time step t we give attention bias value of 5 to knowledge ($b_t^k = 5$) and 1 to dialog history ($b_t^h = 1$), then $\mathbf{b}_t = [5, 5, 5, 1, 1, 1, 1]$. Following this notation, we design three different biasing profiles to explore the extent of control we can achieve from biasing cross-attention[3]:

a) Dialog: for all generation tokens, the cross-attention is always biased towards the dialog history h, i.e., $b_t^k < b_t^h,\ \forall t$.

b) Knowledge: opposite to Dialog profile, here cross-attention is biased towards knowledge k, i.e., $b_t^k > b_t^h,\ \forall t$.

These two profiles mimic a gating strategy between knowledge or dialog history. For the experiments, we instantiate particular profiles where the larger bias is 5x the smaller bias, i.e., $(b_t^k, b_t^h) = (1, 5)$ or $(5, 1)$, for Dialog and Knowledge biasing profiles, respectively.

[2]Although these vectors can be composed of different elements, (i.e at time t the attention biasing factor for i^{th} token of knowledge snippet could be different from that of j^{th} token) we simplify the setup by assigning one attention bias value for knowledge (b_t^k) and another for dialog history (b_t^h) for each generation time step t.

[3]Our biasing profiles are shared across the heads of attention layers. Exploring head-specific biasing is left as a future work.

c) Gradual Knowledge: In contrast to these extreme biasing profiles, we also explore a more practical profile that is motivated from the typical nature of human conversations, where it often is appropriate to start the response by addressing the last utterance of the other party, before introducing new information. In this third profile, decoder cross-attention is initially biased more towards the dialog history, and as the generation time step progresses, the biasing gradually shifts towards the knowledge snippet (see Fig. 4). In particular, through the time steps, knowledge bias value b_t^k increases linearly (with slope s) from 0 up to a certain threshold and dialog bias is kept constant. Similar to earlier, we set $\max\left(b_t^k\right) = 5$, $b_t^h = 1$, and $s = 0.5,\ \forall t$.

To evaluate the responses for the dialog context (k, h), we setup both automatic and human evaluations to measure the responses' *informativeness*, *fluency*, and *relevance*.

Automatic Evaluation. For *informativeness* we use BLEU_k, ROUGE_k, and METEOR_k to compare a generated response with the provided knowledge snippet (k) (instead of ground-truth human response); and for *fluency* we use perplexity with respect to human response (PPL_r). Also for *relevance* we use BERTScore_r (Zhang et al. 2020).

Human Evaluation. For *fluency*, the annotators are asked to make a yes/no decision on the question "*Does the language of the response seem correct?*". To evaluate the *relevance* of responses, we follow prior works (Shin et al. 2019; Rashkin et al. 2019), ask annotators the following question: "*Regardless of its factual correctness, how appropriate is the response to the conversation?*". For *informativeness* we define a new taxonomy (see Table 2) to capture the amount of manifestation of provided knowledge in the response. Given a response, the annotators are asked to assign relevance and informativeness scores on Likert scale of 1-5.

Here, we note that the relevance metric is not as well defined as other metrics. To ensure correct annotations, we deviate from prior work by explicitly adding the phrase "*Regardless of it's factual correctness*" to highlight the annotators that relevance is not associate with factual correctness and could be opinions, personal recollections or similar subjective content. Another issue in relevance metric is the risk of humans assigning high relevance to potentially irrelevant responses like "*I don't know*". In examining the models outputs we do not see such trends in the generated responses.

Level	Taxonomy
1	Does NOT include anything from the provided knowledge and does NOT provide any facts.
2	Does NOT include anything from the provided knowledge but includes some other facts or opinions (made up or not).
3	Includes some words from the provided knowledge, but makes up facts.
4	Indirectly uses provided knowledge, without making up facts.
5	Directly uses provided knowledge, without making up facts.

Table 2: Proposed taxonomy to evaluate *informativeness* in responses. We prioritize knowledge-oriented responses over uninformative responses (Levels 2-5 vs. 1). Within levels 2-5, we prefer responses that adhere to the provided knowledge (4,5) over responses that mention hallucinated facts (2,3).

We utilize Amazon Mechanical Turk as the annotation platform and appoint three annotators per response sample across all model variants. To ensure high quality for annotations, we opt for annotators that are familiar with dialog evaluation and have a high overall performance as Turkers (95% or higher approval rate and more than 5000 approved HITs). We randomly sample 200 dialog instances from the combined test sets of frequent and rare splits in Topical-Chat (100 each), where each instance has the dialog history (h) with five dialog turns and the provided knowledge snippet (k). However, we notice that the top-selected knowledge snippet for a particular dialog context may not always be entirely relevant for the response. This would affect the human evaluations for informativeness as we specifically ask the annotators to prefer responses where facts from the knowledge snippet is manifested in generated response. Thus, we first filter the test sets before sampling the 200 instances. Specifically, we calculate the ROUGE metric between the knowledge snippet and the human response, and only consider the set of dialog contexts that have a higher value than the mean ROUGE value of 0.2[4]. This filtration ensures that the knowledge snippet is relevant to the dialog context and, as a result, can be a good test bed for measuring informativeness.

Inter-Annotator Agreement (IAA). For fluency, relevance, and informativeness, the respective IAA using Krippendorff's alpha are as follows: 0.545, 0.354, 0.373. As relevance and informativeness are scored on a wider scale of 1 to 5, we categorize this 5-scale Likert scale into three bins comprising the values [1,2], [3], and [4,5]. As seen in the IAA values, we achieve high agreements for fluency. For relevance and informativeness, our IAA scores are similar to (Hedayatnia et al. 2020) where the annotations were on a ranking-based format and not Likert-based. It is known in the literature that Likert-based annotations, due to factors like personal bias of annotators, are prone to have lower IAA scores (Van Der Lee et al. 2019). Having said that, we choose this process as it provides a good average value of each model variant (Khashabi et al. 2021).

[4]Mean ROUGE between knowledge snippet and human response over the Topical-Chat training set is 0.2.

Results. Table 1 summarizes the results of applying ATTN. BIAS for controlling the informativeness of generated responses. From the informativeness columns, we can see that using the ATTN. BIAS knob for biasing the cross-attention towards dialog (Dialog profile) causes the automatic metrics (BLEU_k, ROUGE_k, and METEOR_k) to drop, indicating that the provided knowledge is less incorporated in the responses, as expected. On the other hand, when the bias is towards the provided knowledge snippet, we see that these metrics are significantly higher compared to the base model. This trend also appears in the human evaluation, where the informativeness scores are significantly higher. Specifically, we see that using the bias profile "Knowledge", the human evaluation score for informativeness is 3.84, which is statistically significantly larger than the base model's 3.43.

Regarding fluency and relevance, while we see a slight increase in perplexity as the model is biased with different profiles, the human evaluations do not show any statistically significant difference between the models. This indicates that the ATTN. BIAS knob is able to generate more zero-shot informative responses without a negative impact on the fluency and relevance of the responses, hence balancing the trade-off. Table 3 presents example responses from the model with the different biasing profiles showing a varying amount of informativeness in their generations.

Significance Tests. Comparing the mean statistics of the variants, we perform statistical significance tests between all the variant pairs using the Tukey's HSD test (Table 4). We find that compared to the base model (no control knobs applied), none of the controlled variants have fluency or relevance scores that are statistically significant in difference. In contrast, all the knowledge-oriented variants achieve statistically significantly higher informativeness scores. This highlights that the variants are able to improve on informativeness without compromising on fluency and relevance.

Choosing the bias value. We experiment with a range of biasing values, that include $\{1, 2, 5, 10, 50\}$, to see their efficacy towards improving the informativeness of a response. Table 5 demonstrates that in the vicinity of 2 to 10, as the bias value increases, the informativeness of the responses increases without significant degradation to its fluency. However, using an extreme biasing value like 50 practically means only attending to the provided knowledge and ignoring the dialog history. This results in deterioration of fluency of generations (PPL_r = 12.23). As a result, we choose to use the biasing value of 5 in all the experiments and find it to achieve desired improvements in control without hurting fluency and relevance. In the future, we would like to go beyond human priors and automatically set the bias value. One trivial solution for this is to consider the bias as a learnable parameter so that the model can set its value during inference. However, we refrained from using such strategies as it would require training the parameter and hence deviate from the zero-shot framework that is the focus of this work.

Previous Turns:	**A**: Hello - how are you doing today? **B**: Hello, I am doing good. How are you? What do you think of countries having prime ministers? **A**: I'm doing good thanks for asking. I think it's different. What about you? **B**: It is different, I agree. I am not sure how much power they have. It seems like they can do a little more than a president can. **A**: I saw the president of the UK has a cat with a government title of chief mouser.
Knowledge:	*"Broadly speaking, a 'politician' can be anyone seeking political power in any bureaucratic institution."*
Response:	
Dialog Bias	**B**: I wonder how many people are in that position. I think they can be more than the president though. They can do whatever they want.
Knowledge Bias (Informative)	**B**: That is pretty cool. I wonder what kind of job that is. Politicians can be anyone who seeks to achieve political power in any bureaucratic institution.

Table 3: Example of dialog-oriented and knowledge-oriented responses. The responses are generated with the **Dialog** and **Knowledge** bias profiles, respectively.

Base Model vs.	**Fluency**		**Relevance**		**Informativeness**	
	p-value	SSD	p-value	SSD	p-value	SSD
Knowledge	0.90	No	0.79	No	0.001	Yes
Gradual Knowledge	0.90	No	0.90	No	0.001	Yes

Table 4: Comparing variants to the base model for statistically significant mean difference in human evaluation scores as per Tukey's HSD test. SSD refers to Statistically Significant Difference between the models for $p < 0.001$.

Knob			**Test frequent**		**Test rare**	
	b_t^k	b_t^h	**PPL**$_r$	**ROUGE**$_k$	**PPL**$_r$	**ROUGE**$_k$
None	1	1	9.66	0.22	9.88	0.28
ATTN. BIAS	2	1	9.78	0.25	10.05	0.32
	5	1	10.20	0.28	10.59	0.38
	10	1	10.70	0.32	11.18	0.41
	50	1	12.23	0.38	12.90	0.49

Table 5: Effect of varying intensities of biasing for ATTN. BIAS knob. We keep the bias profile to be **Knowledge** (§ 4.2).

4.3 Context Augmentation Experiments

Setup. This section studies the feasibility of CTX. AUG. knob to control attributes that are *not* present in the dialog context. We first discuss the zero-shot generation of *questions* as the main case study. To check if the results generalize for other desired attributes, we further experiment on generating positive sentiment, feedback dialog acts, and fine-grained questions.

Generating questions is an essential skill towards making dialogs more inquisitive and improving their engagingness with the user (See et al. 2019). For applying the CTX. AUG. knob to generate questions, as per § 3.2, we randomly sample 1000 questions from the Topical-Chat training set and use them as *control phrases*. We then generate the *control code* by averaging the control phrases encoded using a pre-trained BART encoder.

Knobs	**Human Eval**		**% of Questions**
	Fluency	**Relevance**	
None	0.86	3.70	29.1%
CTX. AUG.	0.88	3.62	29.3% (↑ 0.2%)
CTX. AUG.+ ATTN. BIAS	0.87	3.56	**35.2%** (↑ 6.1%)

Table 6: Percentage of generated questions (averaged over five runs) using the control knobs, as well as human evaluation of fluency and relevance.

Base Model vs.	**Fluency**		**Relevance**	
	p-value	SSD	p-value	SSD
CTX. AUG.	0.623	No	0.802	No
CTX. AUG.+ATTN. BIAS	0.871	No	0.263	No

Table 7: Statistically significant mean difference (SSD with $p < 0.001$) in human evaluation scores.

We also experiment with combining the CTX. AUG. knob with ATTN. BIAS knob. Regarding attention biasing, unlike § 4.2, the embedding of dialog context $x = (k, h)$ in this section is prepended with the control code c. As a result the new dialog context could be thought of as $x' = (c, x) = (c, k, h)$, with the overall context represenation being $[\mathbb{C} \oplus \text{enc}(x)]$. We use two values b_t^c and b_t^x to define the bias vector $\mathbf{b}_t$ of the ATTN. BIAS knob. Here b_t^c biases the control code and b_t^x biases the original dialog context $x = (k, h)$. We use the profile where $(b_t^c, b_t^x) = (5, 1)$ for all $t < 6$ and $(b_t^c, b_t^x) = (1, 1)$ for $t \geq 6$, which means that the cross-attention is biased towards the control code for the first six decoder time steps[5], while there is no cross-attention biasing for the remaining time steps.

Evaluation. As the initial and ending parts of a dialog typically include greetings and salutations and might not be appropriate to enforce control, we avoid using the complete test set. Instead, we collect a subset of samples from the Topical-Chat test sets by focusing on more central turns in the dialog. We randomly sample 200 dialog contexts (100 from each *frequent* and *rare* splits of the test set and sized equivalently to the human evaluation setup) where each context comprising of five previous dialog turns and use this consolidated test set to evaluate the efficacy of the control knobs. For that, similar to (See et al. 2019), we use "?" as an indicator for questions and count the percentage of responses that contain a question. We also measure *fluency* and *relevance* through human evaluations similar to § 4.2.

Results. Table 6 summarizes the results of biasing the responses towards questions. The first row of this table represents the base case where no biasing knob is applied. From the numbers, one could note that the fluency and relevance, both evaluated by human annotators, do not statistically significantly change as a result of these knobs (Table 7).

In terms of the number of questions generated, we can see that using CTX. AUG. alone does not generate more ques-

[5]Here, 6 is arbitrarily set to instantiate a biasing profile whose attention distribution for the initial generation tokens are biased towards the control code and for the remaining tokens there is no bias. The results also hold for other values near 6, such that the mentioned bias profile structure is maintained.

Previous Turns	**A**: hi how are you? do you play video games? **B**: I'm doing well how are you? Yes I like playing video games, especially on Xbox! **A**: I have not played Xbox or PS. whhich do you prefer? why? **B**: I like xbox more because honestly it's the only system I've had. I remember getting one way back in 2001. **A**: Microsoft gets many things wrong, but ive heard good things about xbox games. i played games as a kid, but now, im not interested. What games do you like?
Knowledge	*"there are controllers created for quadriplegics so they can play XBOX 360 and other consoles."*
Response	
Base Model	**B**: I agree with you. I've heard that they created controllers for quadriplegics so they can play xbox 360 and other consoles.
CTX. AUG.	**B**: I agree with you on that one! I'm not surprised that they created controllers for quadriplegics, I didn't know that.
CTX. AUG.+ATTN. BIAS	**B**: Did you know there are controllers created for quadriplegics so they can play xbox 360 and other consoles?

Table 8: Randomly sampled dialog along with the generations by control knobs from § 4.3. Questions in the turns are highlighted in green.

tions than the base case (row *None*). However, when this knob is combined with the ATTN. BIAS knob, the number of questions generated is larger (6.1% absolute increase). This indicates that while CTX. AUG. knob is *necessary* to guide the model towards the desired attribute, it is not *sufficient*. Combining it with ATTN. BIAS is key for the CTX. AUG. knob to work. We find similar trends for control over dialog acts and sentiment in next section. This result also has implications on how incorporating ATTN. BIAS could help zero- or few-shot prompting for smaller models (Schick and Schütze 2020).

Table 8 presents a dialog instance from the testing set, with responses from different variants detailed above.

Error Analysis. Through these experiments we find that controlling attributes beyond the original context is more challenging compared to the case where the control objective is over the provided context (§ 4.2). In fact, our experiments revealed that although increasing the intensity of the knobs beyond the settings in Table 6 does increase the number of questions, it comes at the cost of lower relevance.

Knobs	**Size of Biasing Set**			
	10	**100**	**1K**	**10K**
None	56.0 ± 1.6	55.6 ± 4.8	58.2 ± 4.2	53.0 ± 4.0
CTX. AUG.	63.4 ± 3.9	59.6 ± 4.5	58.6 ± 5.4	59.4 ± 6.2
CTX. AUG.+ATTN. BIAS	74.2 ± 2.3	65.2 ± 4.2	70.4 ± 3.7	72.2 ± 4.1

Table 9: Effect of number of question control phrases on the number of generated questions out of 200 response turns.

Effect of Number of Control Phrases. Table 9 explores the effect of the number of control phrases on the control quality. Similar to previous results, we see that CTX. AUG. alone is not effective, but when combined with ATTN. BIAS, it shows a significant increase in the number of questions generated. Overall, no discernible pattern could be concluded regarding the impact of the number of control phrases on the number of questions.

Knobs	**# of Questions**	
	Topical-Chat	**SQuAD**
CTX. AUG.	58.6 ± 5.4	56.6 ± 4.4
CTX. AUG.+ATTN. BIAS	**70.4** ± 3.7	**66.6** ± 4.0

Table 10: Comparing control over the number of generated questions (turn-level) between in-domain (Topical-Chat) and out-of-domain (SQuAD) control phrases.

Effect of Source of Control Phrases. Table 10 shows that there is no significant difference between the two sources (Topical-Chat and SQuAD (Rajpurkar, Jia, and Liang 2018)) of control phrases in terms of the final number of generated questions, which suggests that the source of control phrases might *not* be an important factor, particularly for questions. Moreover, this could also be due to the smoothing-out of domain-specific features from the averaging operation in the CTX. AUG. knob.

4.4 Control for Other Attributes

Next, we show that the control knobs introduced in this work could be used for generating responses with other desirable attributes in K-NRG settings. We first look at generating specific dialog acts. For that, we consider the *feedback* dialog act, such as "*Yeah that's right*" or "*That's pretty extreme*", etc., that acknowledges the previous turn in the dialog and improves the overall dialog flow. We also study the ability of the control knobs to control semantic attributes like generating positive responses.

Feedback Responses. To detect the presence of feedback in responses, we utilize an RNN-based classifier trained to annotate the ISO-based Dialog Act Scheme proposed in (Mezza et al. 2018). This scheme contains feedback as an explicit dialog act category. We establish the evaluation reliability of this RNN model by performing human evaluations where it achieves a high F1 score of 0.83. In Table 11, we can see that using feedback control codes helps with generating significantly more responses that are providing feedback for the previous turn (17.7% $\rightarrow$ 22.1%).

Fine-Grained Questions. To evaluate the control over fine-grained questions, we choose different question types (Hedayatnia et al. 2020) that include PropQ (Yes-no questions; e.g. *Do you like it?*), ChoiceQ (Or-questions; e.g. *Or would you go there instead?*), and SetQ (Wh-questions; e.g., *What is your name?*).

For control phrases, we sample the most frequent phrases of these question types from the training set and curate small control sets of these questions' prefixes. For example, for *PropQ*, we curate prefixes that include *"Do you like"*, *"Do you know"*, *"Have you ever"*, *"Are you a"*, etc. By choosing curated phrases, we aim to show that we can achieve control even with a minimal set of control phrases; and also, there is no particular requirement for the control phrases to be well-formed questions. Future work might explore whether the

Knobs	Control Code	Predictions			
		PropQ	SetQ	ChoiceQ	Feedback
None	None	7.5%	2.6%	0.0%	17.7%
CTX. AUG.	*PropQ*	12.1%	3.3%	0.0%	18.1%
CTX. AUG.+ATTN. BIAS		27.2%	3.1%	0.0%	11.8%
CTX. AUG.	*SetQ*	9.8%	3.6%	0.0%	18.1%
CTX. AUG.+ATTN. BIAS		12.4%	6.8%	0.0%	16.3%
CTX. AUG.	*ChoiceQ*	9.5%	3.4%	0.1%	19.1%
CTX. AUG.+ATTN. BIAS		14.2%	4.9%	0.0%	15.0%
CTX. AUG.	*Feedback*	8.4%	2.5%	0.0%	19.1%
CTX. AUG.+ATTN. BIAS		8.7%	2.5%	0.0%	22.1%

Table 11: Comparing control over percentages of dialog acts (calculated per sentence) across 200 response turns when biased by control codes of respective dialog acts (rows). Configurations of ATTN. BIAS and CTX. AUG. knobs are defined in § 4.3.

Knobs	Sentiment	
	p(**positive**$\|y$)	p-value
None	0.561±0.02	-
CTX. AUG.	0.551±0.01	0.4076
CTX. AUG.+ATTN. BIAS	**0.619**±0.03	0.008

Table 12: Sentiment scores (1→positive and 0→negative) averaged over five runs. Configurations of ATTN. BIAS and CTX. AUG. knobs are defined in § 4.3.

control phrases can be automatically generated as done for classification tasks (Shin et al. 2020).

The results are summarized in Table 11. We can see that generating PropQ and SetQ questions could be accomplished by using control knobs. The control knobs, however, fail to generate ChoiceQ questions. One reason for this could be that such questions are quite rare in the training set of the Topical-Chat, and as a result, the model has not learned how to generate them. We see that the model largely adheres to the provided control code in terms of precision of control, i.e., generating as per the control code. It only fails in ChoiceQ, where conditioning with ChoiceQ increases the number of generated PropQ questions. This could be due to the similarity between ChoiceQ and PropQ questions in general.

Sentiment. Finally, we investigate the ability of our knobs to generate more positive responses. For the respective control code, we use control phrases that include *"That's awesome"*, *"That's cool"*, *"Oh that is great"*, *"It's great to"*, and *"It's wonderful to"*. In Table 12, we see that, similar to the previous experiments, when we combine CTX. AUG. with ATTN. BIAS knob, there is statistically significant increase in the average sentiment scores (measured using an off-the-shelf sentiment classifier [6]). This shows that we are able to improve the positivity of the responses using our knobs. Table 13 presents examples for the fine-grained question types, and feedback and sentiment attributes when generated by the CTX. AUG.+ATTN. BIAS model.

5 Conclusion

In this work, we proposed novel approaches to controlling EDT-NLG models based on two control knobs. We found that the transformer models were robust to these knobs, and their biases did not derail the model's generation quality. Building on this observation, we designed manual interventions that provided predictable control on a variety of attributes. Notably, we found that the combination of the proposed knobs leads to the most favorable results. Specifically, applying the CTX. AUG. knob alone would result in little to no control in the generation process. However, when combined with the ATTN. BIAS knob, it would significantly control the generation process. The CTX. AUG. knob could be thought of as an alternative way of prompting generative language models (Brown et al. 2020), whose zero-shot mode previously worked only on models that are enormous in size (Schick and Schütze 2020). Our work is one of the first to bring zero-shot controllability for models that are orders of magnitude smaller than GPT-3 (e.g., BART-base). While in this work, we focused on understanding the feasibility of the control knobs, in the future, we plan to train models that can jointly plan the control goals and generate accordingly.

[6] https://huggingface.co/transformers/quicktour.html

Previous Turns	**A**: Did you know that the University of Iowa's locker room is painted pink? I wonder why? **B**: I think I did hear something about that. I imagine it is an attempt to psych the other team out. **A**: So, it would be in the visiting team's locker room but not their own? **B**: Right. Teams do all kinds of things to bother the competition. I've heard of teams having heated benches in the winter for themselves but not for the visitors. **A**: I would hate a cold bench. Then again, I wouldn't want to be some place that cold or watching football.
Knowledge	*"That the highest score ever in a football game occurred in 1916 when Georgia Tech defeated Cumberland 222-0"*
Response	
PropQ	**B**: Do you know what the highest score ever was in a football game?
SetQ	**B**: What kind of football do you think they play? I bet they score like 222 points
Feedback	**B**: Yeah, that's pretty extreme. I'd rather not go to Georgia Tech. They beat Cumberland 222-0. That's pretty crazy.
Positive	**B**: That's right. Georgia Tech beat Cumberland 222-0 in 1916. It was a good game . They must have been exhausted.

Table 13: Randomly sampled dialog controlled by different fine-grained attributes.

Acknowledgments

We are grateful to Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, and Gokhan Tur for their valuable help in improving the draft and providing baseline codes for this project. We also thank the anonymous reviewers for their constructive comments.

References

Baheti, A.; Ritter, A.; Li, J.; and Dolan, W. B. 2018. Generating More Interesting Responses in Neural Conversation Models with Distributional Constraints. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3970–3980.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Daniluk, M.; Rocktäschel, T.; Welbl, J.; and Riedel, S. 2017. Frustratingly Short Attention Spans in Neural Language Modeling. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Dathathri, S.; Madotto, A.; Lan, J.; Hung, J.; Frank, E.; Molino, P.; Yosinski, J.; and Liu, R. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ghazvininejad, M.; Shi, X.; Priyadarshi, J.; and Knight, K. 2017. Hafez: an interactive poetry generation system. In *Proceedings of ACL 2017, System Demonstrations*, 43–48.

Gopalakrishnan, K.; Hedayatnia, B.; Chen, Q.; Gottardi, A.; Kwatra, S.; Venkatesh, A.; Gabriel, R.; and Hakkani-Tür, D. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, 1891–1895.

Hedayatnia, B.; Gopalakrishnan, K.; Kim, S.; Liu, Y.; Eric, M.; and Hakkani-Tur, D. 2020. Policy-Driven Neural Response Generation for Knowledge-Grounded Dialog Systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, 412–421.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2020. The Curious Case of Neural Text Degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Huang, M.; Zhu, X.; and Gao, J. 2020. Challenges in Building Intelligent Open-domain Dialog Systems. *ACM Trans. Inf. Syst.*, 38(3): 21:1–21:32.

Keskar, N. S.; McCann, B.; Varshney, L. R.; Xiong, C.; and Socher, R. 2019. CTRL: A Conditional Transformer Language Model for Controllable Generation. *CoRR*, abs/1909.05858.

Khashabi, D.; Stanovsky, G.; Bragg, J.; Lourie, N.; Kasai, J.; Choi, Y.; Smith, N. A.; and Weld, D. S. 2021. GENIE: A Leaderboard for Human-in-the-Loop Evaluation of Text Generation. *CoRR*, abs/2101.06561.

Krause, B.; Gotmare, A. D.; McCann, B.; Keskar, N. S.; Joty, S.; Socher, R.; and Rajani, N. F. 2020. Gedi: Generative discriminator guided sequence generation. *arXiv preprint arXiv:2009.06367*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J. R., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 7871–7880. Association for Computational Linguistics.

Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1412–1421.

Madaan, A.; Setlur, A.; Parekh, T.; Póczos, B.; Neubig, G.; Yang, Y.; Salakhutdinov, R.; Black, A. W.; and Prabhumoye, S. 2020. Politeness Transfer: A Tag and Generate Approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1869–1881.

Madotto, A.; Lin, Z.; Bang, Y.; and Fung, P. 2020. The Adapter-Bot: All-In-One Controllable Conversational Model. *CoRR*, abs/2008.12579.

Mezza, S.; Cervone, A.; Stepanov, E. A.; Tortoreto, G.; and Riccardi, G. 2018. ISO-Standard Domain-Independent Dialogue Act Tagging for Conversational Agents. In Bender, E. M.; Derczynski, L.; and Isabelle, P., eds., *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, 3539–3551. Association for Computational Linguistics.

Niu, T.; and Bansal, M. 2018. Polite Dialogue Generation Without Parallel Data. *Trans. Assoc. Comput. Linguistics*, 6: 373–389.

Peng, N.; Ghazvininejad, M.; May, J.; and Knight, K. 2018. Towards Controllable Story Generation. In *Proceedings of the First Workshop on Storytelling*, 43–49. New Orleans, Louisiana: Association for Computational Linguistics.

Prabhumoye, S.; Black, A. W.; and Salakhutdinov, R. 2020. Exploring Controllable Text Generation Techniques. In Scott, D.; Bel, N.; and Zong, C., eds., *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, 1–14. International Committee on Computational Linguistics.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, 784–789. Association for Computational Linguistics.

Rashkin, H.; Reitter, D.; Tomar, G. S.; and Das, D. 2021. Increasing Faithfulness in Knowledge-Grounded Dialogue with Controllable Features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural*

Language Processing (Volume 1: Long Papers), 704–718. Online: Association for Computational Linguistics.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In Korhonen, A.; Traum, D. R.; and Màrquez, L., eds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 5370–5381. Association for Computational Linguistics.

Schick, T.; and Schütze, H. 2020. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. *CoRR*, abs/2009.07118.

See, A.; Roller, S.; Kiela, D.; and Weston, J. 2019. What makes a good conversation? How controllable attributes affect human judgments. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 1702–1723. Association for Computational Linguistics.

Serrano, S.; and Smith, N. A. 2019. Is Attention Interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2931–2951.

Shaw, P.; Uszkoreit, J.; and Vaswani, A. 2018. Self-Attention with Relative Position Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 464–468.

Shin, J.; Xu, P.; Madotto, A.; and Fung, P. 2019. HappyBot: Generating Empathetic Dialogue Responses by Improving User Experience Look-ahead. *CoRR*, abs/1906.08487.

Shin, T.; Razeghi, Y.; Logan IV, R. L.; Wallace, E.; and Singh, S. 2020. Eliciting Knowledge from Language Models Using Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4222–4235.

Smith, E. M.; Gonzalez-Rico, D.; Dinan, E.; and Boureau, Y.-L. 2020. Controlling style in generated dialogue. *arXiv preprint arXiv:2009.10855*.

Snell, J.; Swersky, K.; and Zemel, R. S. 2017. Prototypical Networks for Few-shot Learning. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 4077–4087.

Syed, A. A.; Gaol, F. L.; and Matsuo, T. 2021. A Survey of the State-of-the-Art Models in Neural Abstractive Text Summarization. *IEEE Access*, 9: 13248–13265.

Van Der Lee, C.; Gatt, A.; Van Miltenburg, E.; Wubben, S.; and Krahmer, E. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, 355–368.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vijayakumar, A. K.; Cogswell, M.; Selvaraju, R. R.; Sun, Q.; Lee, S.; Crandall, D.; and Batra, D. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Wu, Z.; Galley, M.; Brockett, C.; Zhang, Y.; Gao, X.; Quirk, C.; Koncel-Kedziorski, R.; Gao, J.; Hajishirzi, H.; Ostendorf, M.; and Dolan, B. 2020. A Controllable Model of Grounded Response Generation. *CoRR*, abs/2005.00613.

Yang, B.; Tu, Z.; Wong, D. F.; Meng, F.; Chao, L. S.; and Zhang, T. 2018. Modeling Localness for Self-Attention Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4449–4458.

Yang, K.; and Klein, D. 2021. FUDGE: Controlled Text Generation With Future Discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3511–3535.

Yang, S.; Wang, Y.; and Chu, X. 2020. A Survey of Deep Learning Techniques for Neural Machine Translation. *CoRR*, abs/2002.07526.

You, W.; Sun, S.; and Iyyer, M. 2020. Hard-Coded Gaussian Attention for Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7689–7700.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, 2204–2213. Association for Computational Linguistics.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.