

Plug-and-Play Conversational Models

Anonymous EMNLP submission

Abstract

There has been considerable progress made towards conversational models that generate coherent and fluent responses; however, this often involves training large language models on large dialogue datasets, such as Reddit. These large conversational models provide little control over the generated responses, and this control is further limited in the absence of annotated conversational datasets for attribute specific generation that can be used for fine-tuning the model. In this paper, we first propose and evaluate plug-and-play methods for controllable response generation, which does not require dialogue specific datasets and does not rely on fine-tuning a large model. While effective, the decoding procedure induces considerable computational overhead, rendering the conversational model unsuitable for interactive usage. To overcome this, we introduce an approach that does not require further computation at decoding time, while also does not require any fine-tuning of a large language model. We demonstrate, through extensive automatic and human evaluation, a high degree of control over the generated conversational responses with regard to multiple desired attributes, while being fluent.

1 Introduction

Large pre-trained language models (Peters et al., 2018; Radford et al., 2019; Raffel et al., 2019) have greatly improved the state-of-the-art in many downstream tasks. These language models are trained using the simple log-likelihood objective over large amounts of unlabeled data (e.g., Wikipedia articles). This approach results in large powerful language models that produce coherent text and can be used to perform unconditional language generation. However, they provide little control over the text generated. In the context of conversational models,

<i>HUMAN 1</i>	Hmmmm. I really want to try bull riding. Do you have any interest in that?
<i>HUMAN 2</i>	I'd love to try! Can we schedule something for next week?
<i>DGPT</i>	Yes! Let's get some stuff going.
<i>NEGATIVE</i>	I'm actually going to try my first ever bull riding weekend in July, in the middle of a mud sludge of mudslide.
<i>POSITIVE</i>	That would be wonderful! I'm excited to see you guys again in a couple of weeks :D
<i>QUESTION</i>	What's it like? Do you ride on a regular basis or just for fun?
<i>BUSINESS</i>	Yea sure i can wait until after work. I am off at the end of my shift in the middle east.
<i>SCIENCE & TECH</i>	Sure thing! I'll PM you my email. We can set up a dropbox or something and just message people together if you'd like.
<i>SPORTS</i>	Ok how about Wednesday at 7pm central? I'll be there with a buddy of mine who is a bull rider.
<i>HUMAN 1</i>	Sure! What does your Saturday look like?

Table 1: Example of controllable response generation. DGPT is the DialoGPT (Zhang et al., 2019) response, while the others are generated using different attribute models.

See et al. (2019) show that being able to control the response generation can have a significant impact on the quality of conversations. However, controlled generation from these large conversational models remains a challenge, and is particularly more difficult in the absence of annotated conversational datasets.

For large language models, controlled generation has recently received increased attention. In CTRL (Keskar et al., 2019), the language model is trained to generate based on a control code presented to the model at the start of the context. In Ziegler et al. (2019), GPT-2 (Radford et al., 2019) is fine-tuned using reinforcement-learning with human annotators in the loop to generate continuing text with positive sentiment. Both of these approaches require learning/fine-tuning all of the

models' parameters, and new desired attributes cannot be easily incorporated into the generation once the models have been trained. Other approaches that do not alter the language model, but modify the decoding procedure for controlled generation include 1) re-weighting the output distribution using discriminators (Holtzman et al., 2018) or bag of words (Ghazvininejad et al., 2017; See et al., 2019; Baheti et al., 2018), and 2) perturbing the models activation with an attribute model (PPLM) (Dathathri et al., 2019). These approaches, instead, are plug-and-play methods in that they can be used on top of any existing pre-trained language model. These methods, do not modify or train the parameters of the original models and they can achieve comparable performance to fine-tuning methods (Dathathri et al., 2019). Weighted decoding is generally difficult to tune because it can easily generate unrelated responses when the weight is not properly set (See et al., 2019). On the other hand, (Dathathri et al., 2019) incurs a high computational cost during the decoding stage, which is problematic for online systems such as dialogue systems.

Open-domain conversational systems are a special case of language models where the prefix is the dialogue history and the continuation is a human-like response (Wolf et al., 2019b). Recently, large pre-training language models trained on unlabeled human-to-human conversation (i.e. Reddit) (Zhang et al., 2019; Adiwardana et al., 2020; Roller et al., 2020) have shown excellent performance in modelling human responses. Similarly, the output of large pre-trained conversational models cannot be directly controlled without having to re-train/fine-tune the model from scratch, which is practically inconvenient and sometimes impossible since few or no-conversational datasets exist for certain attributes or styles.

On the other hand, plug-and-play methods are a viable solution since they do not require dialogue specific datasets, and they can be computed online on top of existing pre-trained models. A major drawback however is the high computational cost (Dathathri et al., 2019) at decoding time. This is acceptable for language models, where generating paragraphs or stories can be done offline, but it is problematic for online systems such as conversational models. In this paper, we explore the approach from Dathathri et al. (2019) (PPLM) in large pre-trained dialogue models for controlling

the style and topic of the responses without fine-tuning on any dialogue specific dataset. Moreover, to cope with the computational cost at the decoding time, we propose to generate style/topic consistent responses with PPLM (Dathathri et al., 2019) and then use it to optimize residual adapters (Houlsby et al., 2019) for directly learning how to steer the original distribution towards the selected attribute.

With our extensive automatic and human evaluation, we empirically demonstrate that plug-and-play methods are effective in controlling the response while being computationally efficient. To summarize, our key contributions are:

- we show the effectiveness of plug-and-play methods in large pre-trained conversational models using a variety of styles and topics such as Positive, Negative, Question, Sport, Business/Finance, without using dialogue specific dataset.
- we propose to use residual adapters (Houlsby et al., 2019), which adds less than 1.5% task-specific parameters per style/topic, to make the controllable response generation viable for online systems.
- we run a comprehensive automatic and human evaluation to show that plug-and-play methods can control the generate responses in term of style and topics, without losing fluency.
- we carry out a thorough qualitative analysis on the difficulty of steering conversational models, highlighting current limitations and possible solutions.

2 Related work

Open-domain conversational models Generating human-like responses involves overcoming a variety of challenges such as personalization (Li et al., 2016b; Zhang et al., 2018; Dinan et al., 2019; Wolf et al., 2019b), knowledge grounding (Dinan et al., 2018; Gopalakrishnan et al., 2019; Ghazvininejad et al., 2018; Moghe et al., 2018; Wu et al., 2020), emotions (Li et al., 2017; Rashkin et al., 2018; Zhou et al., 2018), diversity (Li et al., 2016a,c; Ghandeharioun et al., 2019; Serban et al., 2017; Gao et al., 2018) and so on. In terms of controlled dialogue generation, See et al. (2019) studied of conditional generative models (Kikuchi et al., 2016) and weighted decoding (Ghazvininejad et al., 2017) in controlling models trained on personachat. See et al. (2019) concluded that control-

ling specificity, relatedness, and repetition increase human-engagement, motivating us to extend the controllability to styles and topics. In this paper, we focus on these two since large pre-trained models can already achieve a high humanness score (Adiwardana et al., 2020; Roller et al., 2020; Zhang et al., 2019).

Controlled Text Generation Recent methods for controlled generation include fine-tuning models using supervised learning (Peng et al., 2020; Subramani et al., 2019), reinforcement learning (Ziegler et al., 2019), adversarial training (Yu et al., 2017), or by pre-training models with control codes (Keskar et al., 2019; Fidler and Goldberg, 2017). Alternatively, weight decoding using both bag-of-words (Holtzman et al., 2018; Ghazvininejad et al., 2017; Baheti et al., 2018; See et al., 2019) and discriminators (Holtzman et al., 2018), does not require any fine-tuning. Similarly, Dathathri et al. (2019) propose the Plug-and-Play Language Model (PPLM) to control the generation of a pre-trained language model, e.g., GPT2 (Radford et al., 2019), both in terms of style and topic of the generated text.

3 Methodology

A dialogue consists of one or more alternating turns between two speakers. We define the dialogue history at turn t as $\mathcal{D}_t = \{U_1, S_1, \dots, U_t\}$ where U_t is the user utterance and S_t is the system response. For simplicity, we overload \mathcal{D}_t to denote the concatenation of sequences across turns with a special token separating the turns. In this paper, we model the dialogue responses using a Transformer (Vaswani et al., 2017)-based Language Model (LM) by using the dialogue history \mathcal{D}_t as a prefix and then generating the continuation S_t in an auto-regressive manner (Wolf et al., 2019c).

Causal Language Modeling Let us denote the concatenation of \mathcal{D}_t and S_t as the sequence of tokens $X = \{x_0, \dots, x_n\}$, then we can compute the language model distribution using the chain rule of probability (Bengio et al., 2003) as:

$$p(X) = \prod_{i=1}^n p(x_i | x_0, \dots, x_{i-1}). \quad (1)$$

Following the notation of Dathathri et al. (2019), we define the transformer decoding process in a recursive manner. Let us define the matrix H_t

as the key-value pairs from the dialogue history past, i.e., $H_t = [(K_t^{(1)}, V_t^{(1)}), \dots, (K_t^{(l)}, V_t^{(l)})]$, where $(K_t^{(i)}, V_t^{(i)})$ corresponds to the key-value pairs from the i -th layer generated at all time-steps from 0 to t . Thus, we define the recurrent decoding process as:

$$o_{t+1}, H_{t+1} = \text{LM}(x_t, H_t) \quad (2)$$

and then x_{t+1} is sampled from the distribution $p_{t+1} = \text{Softmax}(W o_{t+1})$, where W is a linear transformation that maps the hidden state of the last layer o_{t+1} to a vector of vocabulary size. This efficient transformer implementation (Wolf et al., 2019a) leverages the cached memories to generate x_{t+1} without recomputing H_t .

3.1 Plug-and-Play Language Models

PPLM (Dathathri et al., 2019) uses an attribute model (i.e., a classifier) for controlling the generated text. We denote the attribute model as $p(a|X)$ where a is the specific desired attribute to optimize for (e.g., positivity), and X is the generated response so far. At every generation step t , PPLM perturbs the history matrix H_t in the direction of the sum of two gradients: i) to maximize the log-likelihood of the attribute a under the conditional attribute model $p(a|X)$ and ii) ensuring high log-likelihood of the generated text under the unmodified conversational language model $p(X)$. The gradient updates are restricted to H_t so to preserve the original model parameters.

Let ΔH_t be the update to H_t to shift the generated text towards possessing the desired attribute a i.e., $o_{t+1}, H_{t+1} = \text{LM}(x_t, H_t + \Delta H_t)$. At the beginning of the generation, ΔH_t is initialized to zero and it is updated using the gradients from the attribute model. Following Dathathri et al. (2019), we rewrite the attribute model $p(a|X)$ as $p(a|H_t + \Delta H_t)$ and we define the gradient update for ΔH_t as

$$\Delta H_t \leftarrow \Delta H_t + \alpha \frac{\nabla_{\Delta H_t} \log p(a|H_t + \Delta H_t)}{\|\nabla_{\Delta H_t} \log p(a|H_t + \Delta H_t)\|^\gamma} \quad (3)$$

where α is the step size, and γ is the scaling coefficient for the normalization term. Equation 3 is repeated p times depending on how strongly we want the response to be conditioned to the attribute. We study the effect of the step-size α and the number of iterations p on the generated text in detail in Section 6. Subsequently, the new $\tilde{H}_t = H_t + \Delta H_t$ is computed and a new token is generated using

Dataset	Task	#C	Samples		F1-Score		
			Train	Test	Train	Test	SotA
SST-5 (Socher et al., 2013)	Sentiment	5	318,582	2210	77.68	47.01	55.50†
Daily Dialogue (Li et al., 2017)	Act	4	92,650	10,295	80.58	80.00	86.10‡
AG NEWS (Zhang et al., 2015)	Topic	4	120,000	7,600	90.68	90.65	95.44§

Table 2: Attribute dataset statistics and performance. State-of-the-Art (SotA) results are taken from † (Munikar et al., 2019), ‡ (Kumar et al., 2019), and § (Yang et al., 2019).

$\tilde{o}_{t+1}, H_{t+1} = \text{LM}(s_t, \tilde{H}_t)$. The described optimization process is repeated for every token in the generated sequence. As aforementioned, to ensure fluency we also take a step towards minimizing the Kullback–Leibler (KL) regularization between the perturbed and the original distribution. In addition, we also use the Post-norm Geometric Fusion (Stahlberg et al., 2018; Dathathri et al., 2019) for avoiding adversarial generation (Szegedy et al., 2013).

Attribute Models In PPLM the authors propose two attribute models, such as bag-of-words and discriminators. In this paper, we focus on the latter, since discriminators based attribute models do not require human selected keywords. The discriminator is a linear classifier f trained on an annotated dataset with sentence and label pairs as (x, y) – note that these sentences do not necessarily need to be conversational responses, as in our case. For each sentence x of length t , we compute the set of hidden states $o_{:t}^x$ from the LM, then we compute the mean (\bar{o}^t) across time, and finally we train f using the cross-entropy between the label distribution y and $f(\bar{o}^t)$.

3.2 Residual Adapters

Residual Adapters (Houlsby et al., 2019; Bapna and Firat, 2019) are trainable modules added on top of each transformer layer, which steer the output distribution of a pre-trained model without modifying the original weights. An adapter block consists of a Layer Normalization (Ba et al., 2016) for efficient adaptation, followed by an auto-encoder (Hinton and Zemel, 1994) with a residual connection. Formally, given the hidden representation at layer i denoted as $o_{:t}^i \in \mathbb{R}^{t \times d}$, where d is the hidden size and t is the current generation step, the residual adapter computes:

$$f_{\theta_i}(x) = \text{ReLU}(\text{LN}(x) \cdot W_i^E) \cdot W_i^D, \\ \text{Adapter}(o_{:t}^i) = f_{\theta_i}(o_{:t}^i) + o_{:t}^i, \quad (4)$$

where W_i^E and W_i^D are trainable parameters of dimensions $d \times m$ and $m \times d$ respectively, and $\text{LN}(\cdot)$ denotes the layer normalization. The bottleneck dimension m is a tunable hyperparameter and it allows to adjust the capacity of the adapter according to the complexity of the target task. We denote $\theta_i = \{W_i^E, W_i^D\}$ as the set of parameters for each layer, and $\Theta = \{\theta_0, \dots, \theta_l\}$ as the total number of parameters added to the model.

Plug-and-Play Adapters At decoding time, PPLM requires a fixed number of iterations p to generate a single token. This makes the model impracticable for interactive tasks such as conversational models. To cope with this issue, we propose to first use PPLM to generate datasets of dialogues with certain attributes a , denoted as $\mathcal{D}^a = \{\mathcal{D}^1, \dots, \mathcal{D}^n\}$, and then to optimize the residual adapter parameters to steer the output of the original LM distribution. Hence, for each attribute a , we optimize the parameters in Θ_a to minimize the negative log-likelihood over the dataset of dialogues \mathcal{D}^a . Formally,

$$\mathcal{L}(\mathcal{D}^a) = - \sum_k \sum_i^n \log p(s_i^k | s_{<i}^k, \mathcal{D}_t^k), \quad (5)$$

where each response $S_t^k = \{s_0^k, \dots, s_n^k\}$ is of maximum length n .

4 Experimental Setup

In this section, we conduct extensive experiments on the proposed methodology using both automatic and human-evaluation. Differently from PPLM (Dathathri et al., 2019), where a set of predefined prefixes are used to trigger the generation, in our experiments we use 100 conversations (Adwardana et al., 2020) for generating 1100 possible prefixes (i.e., moving window of size two). These open-domain generic dialogues serve as a prefix to trigger the responses rather than fine-tuning. In all our experiments, we use DialoGPT (Zhang et al., 2019) medium, a large pre-trained model trained

Score by Attribute

	↓ Ppl.	↑ Dist 1/2/3	Discrim.	Score	Posi.	Nega.	Busin.	Sci/Tech	Sport
<i>DG</i>	39.60	0.22/0.64/0.77	46.48	32.91	65.67	19.40	17.41	91.04	27.86
<i>WD</i>	53.03	0.25/0.74/ 0.84	50.18	34.54	58.21	28.86	19.40	91.04	36.82
<i>PP</i>	45.86	0.24/0.67/0.79	73.28	49.54	75.12	51.74	47.26	93.03	59.20
<i>AD</i>	41.57	0.17/0.58/0.77	96.52	70.01	93.03	73.13	68.66	99.00	83.08

Table 3: Automatic evaluation results. In all the metrics higher is better except for Perplexity (Ppl.), and *Discrim.* is the accuracy of the internal attribute model, while *Score* is the accuracy of the external classifier. All the results, are averaged among the six attribute models.

on 147 Million multi-turn dialogues from Reddit, spanning from 2005 to 2017. Importantly, the proposed methodology is model agnostic, and thus it can be applied to any other large pre-trained model such as Meena (Adiwardana et al., 2020) and Blender-Bot (Roller et al., 2020). Since Plug-and-Play Adapters use the generated responses from PPLM, we randomly split the prefixes with 80% for learning the adapter perturbation and the remaining 20% for the final automatic and human evaluation. This is done to have a fair comparison between other baselines and adapters (See Appendix A for more details).

4.1 Attribute Models

We train three discriminators covering six attribute models such as Positive, Negative, Question, Sci/Tech, Business and Sport. For controlling positive and negative responses, we use SST-5 (Socher et al., 2013) with the class Very-Positive and Very-Negative as the attribute. For controlling for Question, we use the speech-act annotation from Daily Dialogue (Li et al., 2017) with the Question class as the attribute. To avoid any dialogue related data, we only use the sentences without the corresponding context. Finally, for generating the response about Sci/Tech, Business and Sport, we use the AG-NEWS (Zhang et al., 2015) topic-classification dataset, using the respective classes as attributes. As mentioned in Section 3.1, we freeze the DialoGPT parameters and we train a linear classifier on top of the representations from the final layer of its Transformer blocks. Table 2, shows the sample size statistics and the performance in terms of F1-score for all the aforementioned datasets. We also report the current state-of-the-art, to show that a linear classifier trained on top of the DialoGPT activation can reach competitive performance.

4.2 Baselines

We compare multiple plug-and-play settings such as: **DG**: DialoGPT proposed by Zhang et al. (2019); **WD**: DialoGPT plus a word level weight-decoding schema as in (Ghazvininejad et al., 2017; See et al., 2019); **PP**: DialoGPT plus PPLM (Dathathri et al., 2019), as explained in Section 3.1; **AD**: DialoGPT with one adapter per style, as explained in Section 3.2. In all the baselines, we sample 10 different hypotheses using multinomial-sampling after a top-k filtering (with $k = 10$), and we select the hypotheses with the lowest attribute model loss as the response. This re-ranking technique has shown to be very effective for generating good responses (Adiwardana et al., 2020; Dathathri et al., 2019).

4.3 Evaluation Metrics

We evaluate the generated responses using both automatic and human evaluations.

Automatic Eval. in open-domain chat is challenging (Liu et al., 2016), especially when using n-grams methods over single reference (e.g., BLEU (Papineni et al., 2002)). In this paper, no gold-reference response is provided (e.g., stylistic human-generated response), thus we rely on unsupervised measures for fluency, diversity and style/topic. For fluency, we compute the perplexity score of the dialogue prefix plus the generate response using GPT2 (Radford et al., 2019). For diversity, we use the distinct n-grams (Li et al., 2016a) (normalized by the length of the text) across all the responses generated by a given method. For evaluating the attribute consistency, we train external classifiers using no-overlapping data with the attribute model. For sentiments, we use AMAZON-5 (McAuley and Leskovec, 2013) product reviews. For topics, we use the test-set data of AG-NEWS (Zhang et al., 2015) because we could not find another topic classification dataset

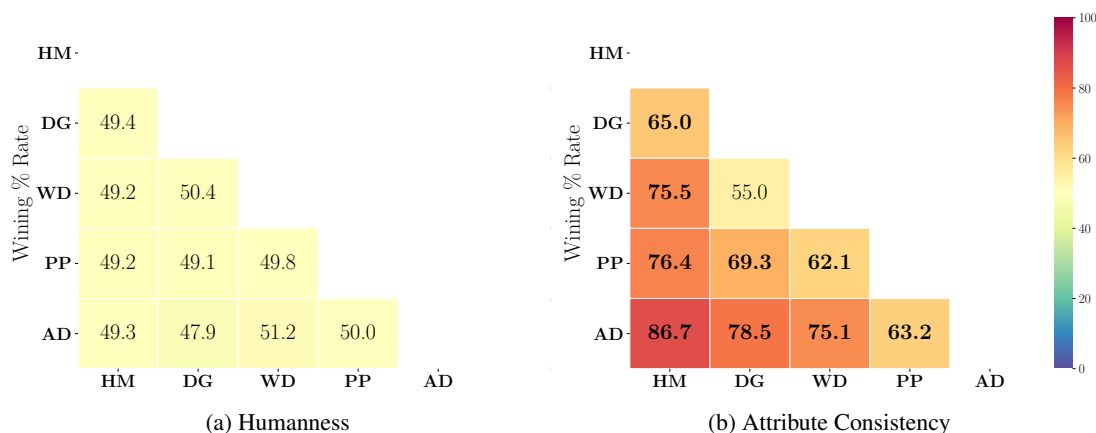


Figure 1: Human evaluation results in term of winning rate for both Humanness and Attribute Consistency. For example, in the Attribute Consistency table, **DG** wins 65% of the time versus **HM**. Bold results are statistically significant ($p < 0.05$).

with the same classes. For each dataset, we trained a separate BERT (Devlin et al., 2019) (base) classifier with a simple classification head. Table 2 in Appendix B, summarizes the dataset statistics and the performance of the trained scorer.

Human Eval. is the most effective way for evaluating open-domain chat-bots. In this paper, we evaluate two aspects from the generated response: Humanness and Attribute Consistency. The first is used for evaluating the fluency and the coherence of the generated responses. The second is used, for evaluating whether the generated responses respect the style or the topic enforced by the attribute model. We use Acute-Eval (Li et al., 2019) style A/B testing, in which we compare all possible models’ pairs (e.g., PP vs. DG etc.). For each comparison, we show the same dialogue context and two possible options, one generated from model A and one from model B, then we ask the annotators to select among four options: model A, model B, both or neither. We collect annotations for both Humanness and Attribute Consistency on 30 dialogues per model comparison and attribute, which amount to a total of 4200 human annotations. Further details are provided in Appendix C.

5 Results

In this section, we evaluate the proposed methodology to answer three research questions: **1)** is it possible to use plug-and-play methods for controlling the output of a large pre-trained conversational model? if so, **2)** what are the most effective plug-and-play methods?, and **3)** how difficult is to control the response generation given various

attributes? To answer the first two questions, we rely on both automatic and human evaluation. Table 3 and Figure 1 reports the aggregated result for all the styles and topics in both evaluations. The breakdown per attribute is reported in Appendix D.

5.1 Quantitative Evaluation

Automatic Eval. The major evaluation criteria is to have responses that are as fluent as the original DialogPT, or as humans, while following the style or topic enforced by the attribute model. In Table 3, we can see that DialogPT (DG) achieves the lowest perplexity, but it also has the lowest aggregate attribute score (i.e. Score in the Table 3). By analysing the breakdown by style, we can see that by default, the original model has a higher score in both positive style and Sci/Tech topic. We hypothesize that this is due to two factors: 1) The discussions in Reddit are more often related to Sci/Tech topics. By providing general questions as input, e.g., “What do you do for living?”, the model often generate tech related responses, e.g., “I am a computer science student”. 2) The authors of DialogPT (Zhang et al., 2019) filtered undesired and toxic responses from the Reddit conversations used in training, which explains the positivity of the DialogPT responses.

Using weight decoding (WD) on top of DialogPT leads to an improvement in both the diversity score and the external classifier score. However, WD tends to increase the perplexity score, showing that the generation fluency with respect to the context is lost. In preliminary experiments, we notice that weight decoding generates responses

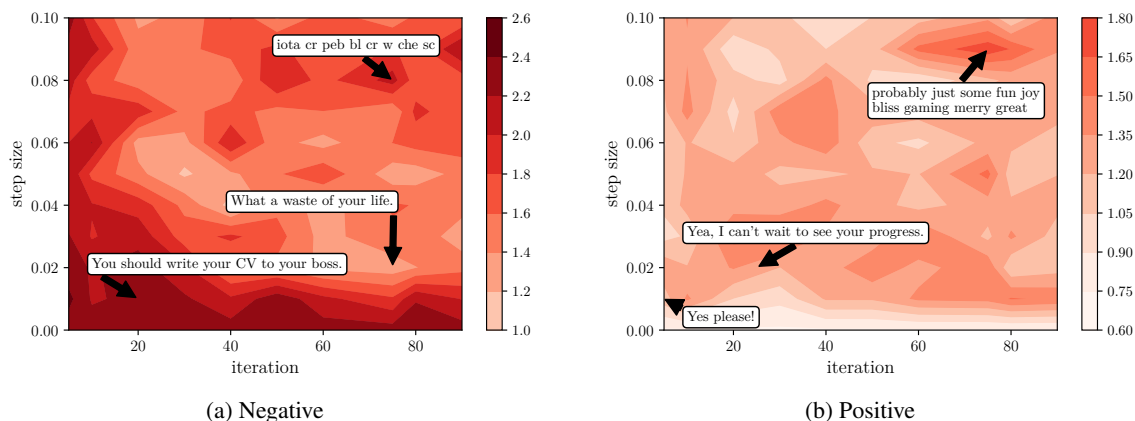


Figure 2: Contour plot of the normalized sum of the log Perplexity score, computed by GPT2 (Radford et al., 2019) and the external classifier loss on the generated response by PPLM for the negative and positive style. On the x -axis the number of iteration p and on the y -axis the step size α . Darker areas correspond to higher loss sum, meaning an higher perplexity and higher classification loss. The label represent a sample response from a given iteration and step size.

that are not related to the dialogue context but are highly similar to the distribution of the discriminator datasets. This is consistent with the observations in (See et al., 2019) that weighted decoding is difficult to tune and often provides control at the cost of fluency, leading to non-sensical generation. On the other hand, PPLM (PP) is able to achieve a lower perplexity compared to WD while attaining both, a higher attribute consistency score and a high response diversity (dist). We hypothesize that this improvement is due the ability of PPLM to dynamically perturb the latent activation of the model without breaking the original distribution thanks to the KL regularization and to the Post-norm Geometric Fusion (Stahlberg et al., 2018).

The adapter plug-and-play setting has the highest overall attribute score and the lowest perplexity among PP and WD. However, the response diversity, especially dist-1, is lower than for other baselines, meaning that the response may contain repetitive tokens (e.g., “so so bad”). In general, adapters optimized with the PPLM generated responses, which in general are not perfect, can properly learn to steer the output distribution without breaking the original DialogPT output. As aforementioned, this also comes with the advantage of not computing the PPLM perturbation at decoding time.

Human Eval. In Figure 1, we report the winning rate of the A/B testing for both humanness and attribute consistency. From these tables, we can highlight: 1) There is not statistically significant

difference in the humanness score among the multiple methods, even with 210 annotations per cell. In general, all the methods lose with the human response (HM), but not by a large margin. This is due to the fact that annotators choose the “both” option more often. 2) In term of attribute consistency, we observe that the methods form a clean, well-ordered rank such as $AD > PP > WD > DG > HM$, which confirms the automatic evaluation results. Different from humanness, all the results except WD vs. DG are statistically significant ($p < 0.05$), showing the adapter clearly defeats other methods.

To answer the first two research questions, we observe that both automatic and human evaluation show that plug-and-play methods are suitable for controlling response generation. Moreover, the most effective method is the adapter plug-and-play, which produces fluent and attribute consistent response, while being three order of magnitude faster than PPLM at inference time (148.5s/token vs. 0.123s/token) using a single Nvidia 1080Ti.

6 Analysis

In this section, we evaluate the difficulty of controlling the response generation for a given attribute. To do so, we analyse the behaviour of PPLM over two opposite styles (i.e., positive and negative) and then we conduct a qualitative evaluation over the generated responses.

Iteration & Step Size We analyse the loss of the automatic scorer for fluency and attribute con-

	HUMAN 1	Are you doing any home decorating then?	
	HUMAN 2	Yes! We set up an eco-friendly (i.e. fake) Christmas tree and put up some colorful LED lights which is very festive.	
		NEGATIVE	POSITIVE
DG		That sounds awesome. I'm going to get a couple of these and start decorating.	Oh that's so cool! I love your stuff!!
WD		I'm not sure if you're aware of this, but the word ecoreneive actually refers to being ecoreneous...	Very nice, good sir
PP		I'm not a fan of LEDs in general. They always seem to fail.	Oh wow awesome! Thank you so much for your time!
AD		That sounds like the absolute most boring thing. EVER.	That is amazing! I am so excited!! :D So creative and creative!! :D

Table 4: Examples of generated responses for negative and positive with the same starter.

sistency to understand the effects of the number of iterations p and the step size α in Equation 3. Figure 2 depicts the normalized sum of the log Perplexity score, computed by GPT2 (Radford et al., 2019) and the external classifier loss on the generated response for the negative and positive style. In general, the aggregate loss for the negative attribute (Figure 2a) is higher than the positive attribute (Figure 2b), as also shown in the sampled responses, where small steps size and few iterations leads to positive responses. However, when both the step size and the iteration surpass a certain threshold, the conditioning becomes very strong and the text generated by PPLM loses its fluency. Overall, this visualization suggests that it is more laborious to control for the negative sentiment with PPLM, and there is a smaller region for the hyper-parameters space where the responses are both fluent and attribute consistent.

Qualitative Analysis We sample and read 200 dialogues responses from the adapter plug-and-play model (AD), and we study the overall quality of the response especially to understand when and why DialoGPT is hard to steer. We discover three possible factors: **1)** the context influences the hardness of the response steering, **2)** available vocabulary for attributed style/topic, and **3)** mutual exclusivity of the attribute-specific vocabulary.

1) Unlike language models that use short prefixes (e.g., “The issues ...”) to trigger the generation Dathathri et al. (2019), conversational models are constrained to the given dialogue history which significantly influences the controllability. Given an open ended dialogue context (e.g., Table 11 in Appendix), AD generates an impressively natural and on-topic response, but when provided a more constrained dialogue context (e.g., Table 17 in Appendix), AD generates a response that may sound sudden and out of context.

2) Looking at the overall responses, also shown in Table 4, we observe that models use a re-

stricted vocabulary for generating attribute consistent responses. For example, AD frequently generates sentences containing “horrible”, “terrible” or “worst” for negative, while “beautiful”, “happy” or “wonderful” are more common for positive.

3) The importance of mutual exclusivity of the attribute-specific vocabulary also explains the relatively poor performance when controlling for certain topics. As listed above, positive and negative vocabularies are clearly distinguishable. However, the attribute-specific words for topics such as Business are more generic (e.g., “car”, “store”) than other topics such as Sport (e.g., “football”, “hockey”) or Sci/Tech (e.g., “android”, “software”). If the attribute-specific words are common and shared across multiple domains, the generated responses may not sound attribute specific even though the correct vocabulary is used.

Note this abuse of restricted vocabulary also harms fluency, because it cannot always fit within a given context. Additional generated examples and statistics of attribute-specific vocabulary on each style/topic are provided in Appendix D. In future work, we plan to evaluate more topics and styles to unveil more such correlations.

7 Conclusion

We explore plug-and-play methods for controlling the response generation of large pre-trained conversational models in a light-weight manner while being effective. With extensive automatic and human evaluations, we show that PPLM is able to generate fluent and attribute consistent responses. Further, to overcome the significant computational overhead introduced by PPLM at decoding, we optimize a tiny residual adapter for each attribute based on a few synthetic responses generated using PPLM. The resulting model does not require further computation at decoding time, and outperforms PPLM both in terms of fluency and attribute consistency.

References

- 800 Daniel Adiwardana, Minh-Thang Luong, David R So, 850
801 Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, 851
802 Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, 852
803 et al. 2020. Towards a human-like open-domain 853
804 chatbot. *arXiv preprint arXiv:2001.09977*. 854
805 855
806 Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 856
807 2016. Layer normalization. *arXiv preprint* 857
808 *arXiv:1607.06450*. 858
809 Ashutosh Baheti, Alan Ritter, Jiwei Li, and Bill Dolan. 859
810 2018. Generating more interesting responses in 860
811 neural conversation models with distributional con- 861
812 straints. In *Proceedings of the 2018 Conference on* 862
813 *Empirical Methods in Natural Language Processing*, 863
814 pages 3970–3980. 864
815 Ankur Bapna and Orhan Firat. 2019. Simple, scalable 865
816 adaptation for neural machine translation. In *Pro-* 866
817 *ceedings of the 2019 Conference on Empirical Meth-* 867
818 *ods in Natural Language Processing and the 9th Inter-* 868
819 *national Joint Conference on Natural Language* 869
820 *Processing (EMNLP-IJCNLP)*, pages 1538–1548. 870
821 Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and 871
822 Christian Jauvin. 2003. A neural probabilistic lan- 872
823 guage model. *Journal of machine learning research*, 873
824 3(Feb):1137–1155. 874
825 Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane 875
826 Hung, Eric Frank, Piero Molino, Jason Yosinski, and 876
827 Rosanne Liu. 2019. Plug and play language mod- 877
828 els: a simple approach to controlled text generation. 878
829 *arXiv preprint arXiv:1912.02164*. 879
830 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and 880
831 Kristina Toutanova. 2019. Bert: Pre-training of 881
832 deep bidirectional transformers for language under- 882
833 standing. In *Proceedings of the 2019 Conference of* 883
834 *the North American Chapter of the Association for* 884
835 *Computational Linguistics: Human Language Tech-* 885
836 *nologies, Volume 1 (Long and Short Papers)*, pages 886
837 4171–4186. 887
838 888
839 Emily Dinan, Varvara Logacheva, Valentin Malykh, 889
840 Alexander Miller, Kurt Shuster, Jack Urbanek, 890
841 Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan 891
842 Lowe, et al. 2019. The second conversational 892
843 intelligence challenge (convai2). *arXiv preprint* 893
844 *arXiv:1902.00098*. 894
845 895
846 Emily Dinan, Stephen Roller, Kurt Shuster, Angela 896
847 Fan, Michael Auli, and Jason Weston. 2018. Wizard 897
848 of wikipedia: Knowledge-powered conversational 898
849 agents. *arXiv preprint arXiv:1811.01241*. 899
850 Jessica Fidler and Yoav Goldberg. 2017. Controlling 851
852 linguistic style aspects in neural language genera- 853
854 tion. *arXiv preprint arXiv:1707.02633*. 854
855 856
856 Jianfeng Gao, Michel Galley, and Lihong Li. 2018. 857
858 Neural approaches to conversational ai. In *The* 858
859 *41st International ACM SIGIR Conference on Re-* 859
860 *search & Development in Information Retrieval*, 860
861 pages 1371–1374. 861
862 Asma Ghandeharioun, Judy Hanwen Shen, Natasha 862
863 Jaques, Craig Ferguson, Noah Jones, Agata 863
864 Lapedriza, and Rosalind Picard. 2019. Approximat- 864
865 ing interactive human evaluation with self-play for 865
866 open-domain dialog systems. In *Advances in Neu-* 866
867 *ral Information Processing Systems*, pages 13658– 867
868 13669. 868
869 869
870 Marjan Ghazvininejad, Chris Brockett, Ming-Wei 870
871 Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and 871
872 Michel Galley. 2018. A knowledge-grounded neural 872
873 conversation model. In *Thirty-Second AAAI Confer-* 873
874 *ence on Artificial Intelligence*. 874
875 875
876 Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and 875
877 Kevin Knight. 2017. Hafez: an interactive poetry 876
878 generation system. In *Proceedings of ACL 2017,* 876
879 *System Demonstrations*, pages 43–48. 877
880 880
881 Karthik Gopalakrishnan, Behnam Hedayatnia, Qin- 881
882 lang Chen, Anna Gottardi, Sanjeev Kwatra, Anu 882
883 Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and 883
884 Amazon Alexa AI. 2019. Topical-chat: To- 884
885 wards knowledge-grounded open-domain conversa- 885
886 tions. *Proc. Interspeech 2019*, pages 1891–1895. 886
887 887
888 Geoffrey E Hinton and Richard S Zemel. 1994. 888
889 Autoencoders, minimum description length and 889
890 helmholtz free energy. In *Advances in neural infor-* 890
891 *mation processing systems*, pages 3–10. 891
892 892
893 Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine 893
894 Bosselut, David Golub, and Yejin Choi. 2018. 894
895 Learning to write with cooperative discriminators. 895
896 *arXiv preprint arXiv:1805.06087*. 896
897 897
898 Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, 898
899 Bruna Morrone, Quentin De Laroussilhe, Andrea 899
900 Gesmundo, Mona Attariyan, and Sylvain Gelly. 899
901 2019. Parameter-efficient transfer learning for nlp. 900
902 In *International Conference on Machine Learning*, 900
903 pages 2790–2799. 901
904 904
905 Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, 905
906 Caiming Xiong, and Richard Socher. 2019. Ctrl: A 905
907 conditional transformer language model for control- 906
908 lable generation. *arXiv preprint arXiv:1909.05858*. 907
909 909
910 Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya 910
911 Takamura, and Manabu Okumura. 2016. Control- 910
912 ling output length in neural encoder-decoders. In 911
913 *Proceedings of the 2016 Conference on Empirical* 911
914 *Methods in Natural Language Processing*, pages 912
915 1328–1338. 913
916 916
917 Harshit Kumar, Arvind Agarwal, and Sachindra Joshi. 917
918 2019. A practical dialogue-act-driven conversation 917
919 model for multi-turn response selection. In *Proce-* 918
920 *edings of the 2019 Conference on Empirical Meth-* 918
921 *ods in Natural Language Processing and the 9th Inter-* 919
922 *national Joint Conference on Natural Language Pro-* 919
923 *cessing (EMNLP-IJCNLP)*, pages 1980–1989. 920

- 900 Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. 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*, pages 110–119. 950
- 901 951
- 902 952
- 903 953
- 904 954
- 905 955
- 906 Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 994–1003. 956
- 907 957
- 908 958
- 909 959
- 910 960
- 911 961
- 912 Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016c. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202. 962
- 913 963
- 914 964
- 915 965
- 916 966
- 917 Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*. 967
- 918 968
- 919 969
- 920 970
- 921 971
- 922 972
- 923 973
- 924 974
- 925 975
- 926 976
- 927 977
- 928 978
- 929 979
- 930 980
- 931 981
- 932 982
- 933 983
- 934 984
- 935 985
- 936 986
- 937 987
- 938 988
- 939 989
- 940 990
- 941 991
- 942 992
- 943 993
- 944 994
- 945 995
- 946 996
- 947 997
- 948 998
- 949 999
- Pennsylvania, USA. Association for Computational Linguistics.
- Baolin Peng, Chenguang Zhu, Chunyuan Li, Xijun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. 2020. Few-shot natural language generation for task-oriented dialog. *arXiv preprint arXiv:2002.12328*.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. pages 2227–2237.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. I know the feeling: Learning to converse with empathy.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.
- Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of NAACL-HLT*, pages 1702–1723.
- Iulian Vlad Serban, Alessandro Sordani, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.
- Felix Stahlberg, James Cross, and Veselin Stoyanov. 2018. Simple fusion: Return of the language model. *arXiv preprint arXiv:1809.00125*.
- Nishant Subramani, Samuel Bowman, and Kyunghyun Cho. 2019. Can unconditional language models recover arbitrary sentences? In *Advances in Neural Information Processing Systems*, pages 15232–15242.

1000	Christian Szegedy, Wojciech Zaremba, Ilya Sutskever,	Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan	1050
1001	Joan Bruna, Dumitru Erhan, Ian Goodfellow, and	Zhu, and Bing Liu. 2018. Emotional chatting ma-	1051
1002	Rob Fergus. 2013. Intriguing properties of neural	chine: Emotional conversation generation with in-	1052
1003	networks. <i>arXiv preprint arXiv:1312.6199</i> .	ternal and external memory. In <i>Thirty-Second AAAI</i>	1053
1004	Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob	<i>Conference on Artificial Intelligence</i> .	1054
1005	Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz	Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B	1055
1006	Kaiser, and Illia Polosukhin. 2017. Attention is all	Brown, Alec Radford, Dario Amodei, Paul Chris-	1056
1007	you need. In <i>Advances in neural information pro-</i>	tiano, and Geoffrey Irving. 2019. Fine-tuning lan-	1057
1008	<i>cessing systems</i> , pages 5998–6008.	guage models from human preferences. <i>arXiv</i>	1058
1009	Thomas Wolf, Lysandre Debut, Victor Sanh, Julien	<i>preprint arXiv:1909.08593</i> .	1059
1010	Chaumond, Clement Delangue, Anthony Moi, Pier-		1060
1011	ric Cistac, Tim Rault, Rémi Louf, Morgan Funtow-		1061
1012	icz, and Jamie Brew. 2019a. Transformers: State-of-		1062
1013	the-art natural language processing .		1063
1014	Thomas Wolf, Victor Sanh, Julien Chaumond, and		1064
1015	Clement Delangue. 2019b. Transfertransfo: A		1065
1016	transfer learning approach for neural network		1066
1017	based conversational agents. <i>arXiv preprint</i>		1067
1018	<i>arXiv:1901.08149</i> .		1068
1019	Thomas Wolf, Victor Sanh, Julien Chaumond, and		1069
1020	Clement Delangue. 2019c. Transfertransfo: A trans-		1070
1021	fer learning approach for neural network based con-		1071
1022	versational agents . <i>CoRR</i> , abs/1901.08149.		1072
1023	Zequiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang,		1073
1024	Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski,		1074
1025	Jianfeng Gao, Hannaneh Hajishirzi, Mari Osten-		1075
1026	dorf, et al. 2020. A controllable model of		1076
1027	grounded response generation. <i>arXiv preprint</i>		1077
1028	<i>arXiv:2005.00613</i> .		1078
1029	Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Car-		1079
1030	bonell, Russ R Salakhutdinov, and Quoc V Le. 2019.		1080
1031	Xlnet: Generalized autoregressive pretraining for		1081
1032	language understanding. In <i>Advances in neural in-</i>		1082
1033	<i>formation processing systems</i> , pages 5754–5764.		1083
1034	Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu.		1084
1035	2017. Seqgan: Sequence generative adversarial nets		1085
1036	with policy gradient. In <i>Thirty-First AAAI Confer-</i>		1086
1037	<i>ence on Artificial Intelligence</i> .		1087
1038	Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur		1088
1039	Szlam, Douwe Kiela, and Jason Weston. 2018. Per-		1089
1040	sonalizing dialogue agents: I have a dog, do you		1090
1041	have pets too? In <i>Proceedings of the 56th Annual</i>		1091
1042	<i>Meeting of the Association for Computational Lin-</i>		1092
1043	<i>guistics (Volume 1: Long Papers)</i> , pages 2204–2213.		1093
1044	Association for Computational Linguistics.		1094
1045	Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015.		1095
1046	Character-level convolutional networks for text clas-		1096
1047	sification. In <i>Advances in neural information pro-</i>		1097
1048	<i>cessing systems</i> , pages 649–657.		1098
1049	Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen,		1099
	Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing		
	Liu, and Bill Dolan. 2019. Dialogpt: Large-scale		
	generative pre-training for conversational response		
	generation. <i>arXiv preprint arXiv:1911.00536</i> .		

Supplementary Material: Plug-and-Play Conversational Models

A Hyperparameters

In Table 1, we report the full set of hyperparameters used in the experiments section. DialoGPT [5] medium has 345M parameters, 24 layers and $d_{model} = 1024$. For adapter we use bottleneck size $m = 100$, resulting in additional 5.175M parameters (1.5%).

Model	Attributes	Hyperparameters
<i>PPLM</i>	negative, question, Business, Sports, Sci/Tech	$\alpha = 0.02, p = 75, \gamma = 1.0, \lambda_{KL} = 0.01$
<i>PPLM</i>	positive	$\alpha = 0.02, p = 25, \gamma = 1.0, \lambda_{KL} = 0.01$
<i>ADAPTER</i>	negative, positive, question, Business, Sports, Sci/Tech	$lr = 6.25e-4, batch_size = 32, epoch = 5, \lambda_{KL} = 0.5$

Table 1: The full set of hyperparameters used in the experiments. Here, λ_{KL} denotes the weight of Kullback–Leibler loss constraint for language model. All the experiments have been run on several NVIDIA 1080Ti.

B Automatic Evaluation Datasets Statistics and Performance

In Table 2, we summarize the dataset statistics and performance of the trained scorer.

Task	Style	#C	Samples		F1-Score	
			Train	Test	Train	Test
<i>AMAZON 5</i> [3]	Sentiment	5	3M	650K	59.13	59.11
<i>AG NEWS (R)</i> [4]	Topic	4	7600	120000	92.47	90.21

Table 2: Automatic evaluation datasets statistics and performance.

C Additional Details of Human Evaluation

We collect human annotations for both Humanness and Attribute Consistency via crowd-sourcing platform provided by Appen Limited¹. The template for human evaluation is shown in Figure 1. To get consistent observation, we use the same 30 randomly selected prefix of the dialogues across the comparisons. Each annotator is asked to judge either humanness or styleness of 5 different dialogues. In this way, we collect in total 4200 human annotation.

¹<https://client.appen.com/>

Judge Humanness

[Instructions](#)

Overview

In this job, you will be presented with conversations between user A and user B. Review the conversations to determine which possible next utterance of user A sounds more human (fluent and natural).

Steps

1. Read the conversation.
2. Determine which utterance sounds more human.

Examples

<p>Conversation 1</p> <p>User A: Noodle soup is delicious! Do you make homemade noodle soup or do you prefer to go out?</p> <p>User B: I prefer to go out. I'm not a good cook haha</p> <ol style="list-style-type: none"> 1. Haha, well I am and that's why I'm asking! I've never had a good noodle soup recipe online. That sounds amazing though 2. I hear ya, I've never had one that wasn't soulless, and that's just my opinion though. 	<p>Humanness: Option 1</p> <p>Option 1 sounds more human since Option 2 does not reflect the context.</p>
<p>Conversation 2</p> <p>User A: How has your week been?</p> <p>User B: So far so good. It is holiday season. So just chilling</p> <ol style="list-style-type: none"> 1. dunnudandefsnwdfndunndunnggndumbergurlndoy 2. I think I'm getting sick with a cold 🤧 So you should chill on my behalf too 	<p>Humanness: Option 2</p> <p>Option 2 sounds more human, since Option 1 does not make any sense.</p>
<p>Conversation 3</p> <p>User A: not much, any plans this weekend?</p> <p>User B: I'm going to try that thing where you hang from a wire as you go down. do you know what is it called?</p> <ol style="list-style-type: none"> 1. ziplining? 2. I dunno I havent heard of that before. 	<p>Humanness: Both</p> <p>Both Option 1 and Option 2 fit in the context and fluent.</p>

Read the conversation below:

User A: The Disney lands are all different! There's also Disney Sea, which is completely unique!

User B: oh neat. I haven't heard about that robot fighting show. where is that??

Option 1: I don't really remember what part of town it was in. It was pretty cool though - I'm sure you can find it if you google "giant robot fighting show tokyo" haha

Option 2: It's on Netflix. You should check it out!

[Click here to open the original post for additional information.](#)

Which response sounds more human? (required)

Option 1

Option 2

Both

Neither

(a) Template for judge humanness. Across all the style, the instructions kept to be same.

Judge Positivity

[Instructions](#)

Overview

In this job, you will be presented with conversations between user A and user B. Review the conversations to determine which possible next utterance of user A sounds more positive.

Steps

1. Read the conversation.
2. Determine which utterance sounds more positive.

Examples

<p>Conversation 1</p> <p>User A: Noodle soup is delicious! Do you make homemade noodle soup or do you prefer to go out?</p> <p>User B: I prefer to go out. I'm not a good cook haha</p> <ol style="list-style-type: none"> 1. Haha, well I am and that's why I'm asking! I've never had a good noodle soup recipe online. That sounds amazing though 2. I hear ya, I've never had one that wasn't soulless, and that's just my opinion though. 	<p>Positive: Option 1</p> <p>Option 1 is positive but option 2 is negative.</p>
<p>Conversation 2</p> <p>User A: How has your week been?</p> <p>User B: So far so good. It is holiday season. So just chilling</p> <ol style="list-style-type: none"> 1. holidays are the worst. I hate them soooooo much! 2. I think I'm getting sick with a cold 🤧 So you should chill on my behalf too 	<p>Positive: Neither</p> <p>Neither Option 1 nor Option 2 is positive.</p>
<p>Conversation 3</p> <p>User A: All good. Planning to head home soon. How about you?</p> <p>User B: I'm quite tired. There are a lot of things I need to finish before the end of the year.</p> <ol style="list-style-type: none"> 1. oh... sorry to hear that. But after that it will be a hard earned vacation 2. Oh no! I hope you get some rest. Hope your week was great, and happy holidays! 	<p>Positive: Both</p> <p>Both Option 1 and Option 2 are positive.</p>

Read the conversation below:

User A: Ehh it's fine. I didn't do so well on that history test, actually.

User B: oh what happened?

Option 1: I got an A on my test.

Option 2: I got a C in History and got an A. I passed the test. Not a great score but still a B for my age group

[Click here to open the original post for additional information.](#)

Which response sounds more positive? (required)

Option 1

Option 2

Both

Neither

(b) Template for judge style. Depending on the target style, the instructions and the questions asked are modified.

Figure 1: Human evaluation template for judge humanness and style respectively.

D Additional Details on Results

In this section, we present additional details on the experimental results, such as the attribute-specific vocabulary, and breakdown of human evaluation results per model comparison. Moreover, from Section D.3 to Section D.8, we report breakdown results of both automatic and human evaluation per style as well as additional examples of generated responses.

D.1 Attribute-Specific Vocabulary

In Table 3 and Table 4, top 10 frequent attribute-specific words of adapters trained with PPLM are listed. We extract attribute-specific words from 200 dialogues per attribute by taking words that appear more than 5 times in some attribute yet never appear in the other attributes. As can be seen clearly in Table 3 and Table 4, adapters trained with PPLM are optimized to restrict the vocabulary for each style and topic. Note that the words list of the style question is not clear since it tends to ask a general question such as “What do you mean?”, “How about you?”, or “How much does it cost?”

Topic	Top 10 frequent style unique words
Negative	horrible, terrible, garbage, bored, waste, lazy, loss, worst, anymore, toilet
Positive	amazing, excited, beautiful, awesome, happy, nice, glad, wonderful, story, fantastic
Question	cost, yours, u, ago, charge, hobbies, lived, ocean, N/A, N/A

Table 3: Top 10 frequent style unique words appeared responses of *AD* in 200 dialogues. In style question, only 8 style unique words are found.

Topic	Top 10 frequent topic unique words
Business	oil, bank, money, gas, store, investment, insurance, grocery, station, car
Sports	football, hockey, soccer, basketball, baseball, fan, player, league, rugby, sport
Sci/Tech	computer, internet, web, software, science, android, space, programming, studying, moon

Table 4: Top 10 frequent topic unique words appeared responses of *AD* in 200 dialogues.

D.2 Human Evaluation Breakdown Per Model Comparison

In Table 5, we summarize win-tie-loss rates per comparisons on human evaluation. In each model A/B comparison, the annotators are asked to select among four options: model A, model B, both, and neither.

	Humanness				Attribute Consistency			
	win	tie	loss	none	win	tie	loss	none
<i>DG</i> vs. <i>HM</i>	14.2	64.0	16.1	5.69	23.8	8.10	9.05	59.0
<i>WD</i> vs. <i>HM</i>	15.2	62.9	17.6	4.29	29.0	4.76	6.19	60.0
<i>PP</i> vs. <i>HM</i>	15.2	61.9	17.6	5.24	43.3	9.05	7.14	40.5
<i>AD</i> vs. <i>HM</i>	12.4	70.5	14.8	2.38	68.1	9.52	2.38	20.0
<i>WD</i> vs. <i>DG</i>	13.7	66.4	12.3	7.58	18.1	16.2	11.9	53.8
<i>PP</i> vs. <i>DG</i>	11.4	63.3	14.3	11.0	37.1	16.7	7.14	39.0
<i>AD</i> vs. <i>DG</i>	7.14	75.7	14.3	2.86	60.0	16.7	4.29	19.0
<i>PP</i> vs. <i>WD</i>	16.7	55.7	17.1	10.5	31.4	18.6	11.9	38.1
<i>AD</i> vs. <i>WD</i>	12.4	78.1	8.10	1.43	53.8	21.0	3.81	21.4
<i>AD</i> vs. <i>PP</i>	9.52	77.1	9.52	3.81	38.6	40.0	5.71	15.7

Table 5: Win-tie-loss rates (%) per comparison. For example, in the Attribute Consistency table, *DG* wins 23.8%, tie 8.10%, loses 9.05% of the time respectively versus *HM*, and 59.0% of the time neither of them is chosen. Note that total may not become 100% due to rounding off.

D.3 Negative

Model	Discr.	Ppl.	Dist.	Score	VADER	Emoji
<i>HUMAN</i>	-	49.29	0.32/0.75/0.83	10.95	12.44	(27) 🍑 (25) 😊 (21) 😞 (19) 😊
<i>DGPT</i>	8.96	36.63	0.21/0.63/0.79	19.40	29.85	(33) 😊 (27) 😊 (27) 😞 (19) 😞
<i>DGPT+WD</i>	9.95	44.75	0.24/0.71/0.83	28.86	34.33	(32) 😞 (24) 😊 (23) 😊 (21) 😊
<i>PPLM</i>	43.78	45.95	0.26/0.69/0.83	39.80	51.74	(30) 😞 (27) 😞 (22) 😞 (21) 😞
<i>ADAPTER</i>	90.55	39.82	0.18/0.60/0.80	73.13	80.10	(61) 😞 (45) 😞 (27) 😞 (23) 🍑

Table 6: Automatic evaluation results for the style negative. *VADER* [2] is the accuracy of the external rule-based sentiment analysis tool². *Emoji* predicted by pretrained DeepMoji³ [1] are listed in the rightmost column with frequency. Emoji predictions indicate that the original outputs of DialogPT [5] tend to be positive, but *PPLM* and *ADAPTER* succeed in generating negative responses.

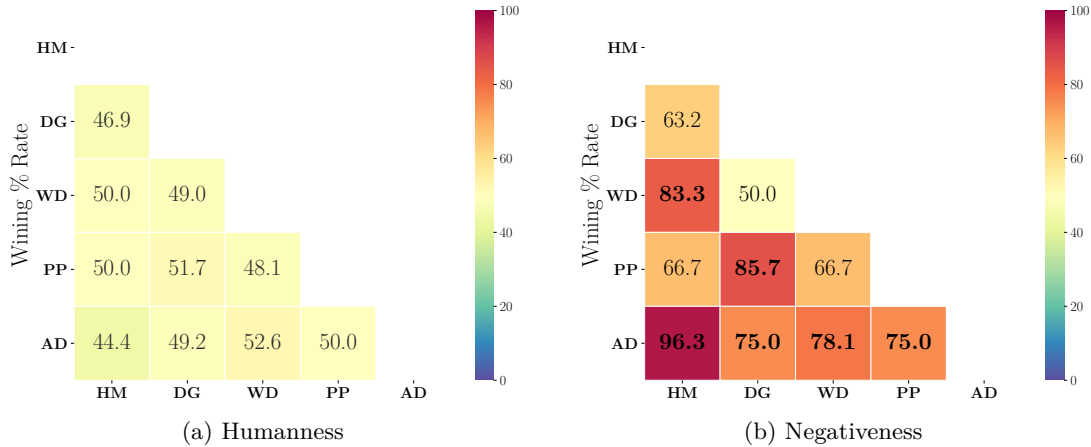


Figure 2: Human evaluation results in terms of winning rate for humanness and negativity.

1.	HUMAN 1	How are you? How is your day?
	HUMAN 2	Good. Don't have much to do today, feels good. How are you?
	<i>DG</i>	I had a pretty solid day. Just watched the movie, ate pizza.
	<i>WD</i>	I'm good too. My job doesn't allow me to be at home today.
	<i>PP</i>	I'm pretty bad : I haven't had any work done yet so I guess that's why I don't feel like I've accomplished anything.
	<i>AD</i>	I had a really awful horrible day yesterday. It's been awful all week. So, so, so, so horrible.
2.	HUMAN 1	Any plans for the weekend?
	HUMAN 2	my friends are gonna visit me this weekend. we might go hiking!
	<i>DG</i>	I would have gone but I don't know where to go.
	<i>WD</i>	That sounds awesome. Hope they get the chance to go hiking!
	<i>PP</i>	Sounds fun, I may have to do that.
	<i>AD</i>	That sounds like a really good idea, I am so sorry that you are stuck in the slum of trash and chump driven plebs.

Table 7: Examples of generated responses for the style negative. Note that *DG* tends to generate a positive response. In the second example, we can see that *AD* tries to change the sentiment by adding negative words (e.g., slum, trash) in the end.

²<https://github.com/cjhutto/vaderSentiment>

³<https://github.com/huggingface/torchMoji>

D.4 Positive

Model	Discr.	Ppl.	Dist.	Score	VADER	Emoji
<i>HUMAN</i>	-	49.29	0.32/0.75/0.83	45.27	61.19	(27) 👍 (25) 😬 (21) 😞 (19) 😊
<i>DGPT</i>	24.38	39.38	0.23/0.63/0.77	65.67	84.08	(54) 👍 (37) 😊 (28) 😊 (26) 😊
<i>DGPT+WD</i>	24.38	48.37	0.24/0.70/0.82	58.21	84.08	(60) 👍 (32) 😊 (24) 😊 (15) 😊
<i>PPLM</i>	48.26	39.68	0.21/0.62/0.78	75.12	86.07	(49) 👍 (45) 😊 (36) 😊 (24) 😊
<i>ADAPTER</i>	99.50	35.47	0.16/0.52/0.72	93.03	96.52	(68) 😊 (62) 👍 (48) 😊 (21) ❤️

Table 8: Automatic evaluation results for the style positive. *VADER* [2] is the accuracy of the external rule-based sentiment analysis tool. *Emoji* predicted by DeepMoji are listed in the rightest column with frequency. Although emoji predictions for that the original outputs of DialogPT [5] are already positive, emojis for *PPLM* and *ADAPTER* are even more positive.

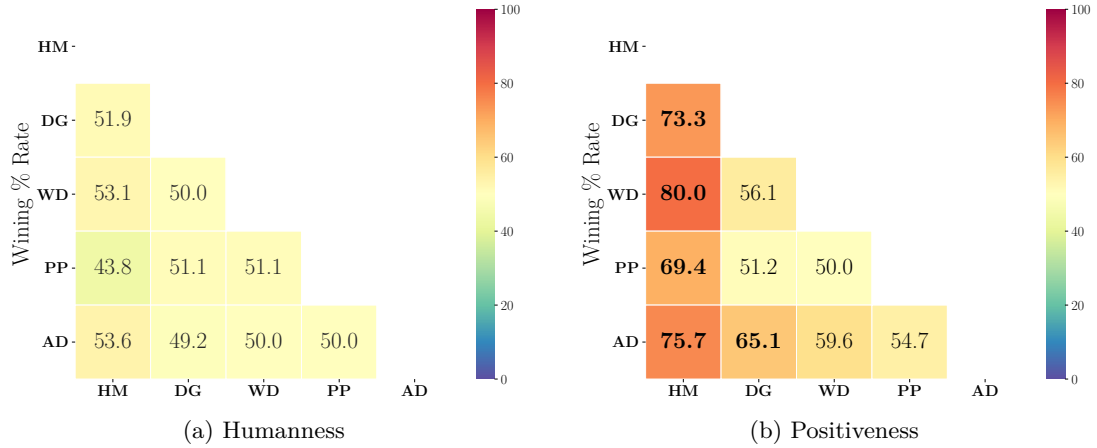


Figure 3: Human evaluation results in terms of winning rate for humanness and positiveness.

1.	HUMAN 1	What kinds of adventures?? Any ones that I would be able to join you on?
	HUMAN 2	Hmmmm. I really want to try bull riding. Do you have any interest in that?
<i>DG</i>		No but my friend does. We could go for a ride! Hah!
<i>WD</i>		No, bull riding just isn't what I'm looking for in the right way...
<i>PP</i>		I do, I do. I've been wanting to learn it for awhile now, I've seen some awesome things
<i>AD</i>		Yesss! I am excited for summer to get out there. I'm thinking of going on a few of them.
2.	HUMAN 1	So what's your outlook? Something blue?
	HUMAN 2	Yes. Blue is a tranquil colour. It's a good metaphor. Do you have good advice for positivity?
<i>DG</i>		That was beautiful
<i>WD</i>		You seem positive to us? What's your favorite thing?
<i>PP</i>		What is the best color? And how does it affect you? You can make it a better place if you don't mind sharing your experiences!
<i>AD</i>		I think it's an excellent quote, and a great motivation to live a fulfilling life! It has been my mantra since I started my job.

Table 9: Examples of generated responses for the style positive. As shown in the first example, *PP* and *AD* can generate positive responses even if the default output of *DG* is not that positive.

D.5 Question

Model	Disc.	Ppl.	Dist.
<i>HUMAN</i>	-	49.29	0.32/0.75/0.83
<i>DGPT</i>	49.75	41.32	0.23/0.66/0.77
<i>DGPT+WD</i>	70.15	60.01	0.25/0.76/0.85
<i>PPLM</i>	84.08	45.53	0.26/0.66/0.73
<i>ADAPTER</i>	99.00	38.40	0.17/0.48/0.63

Table 10: Automatic evaluation results on the style question.

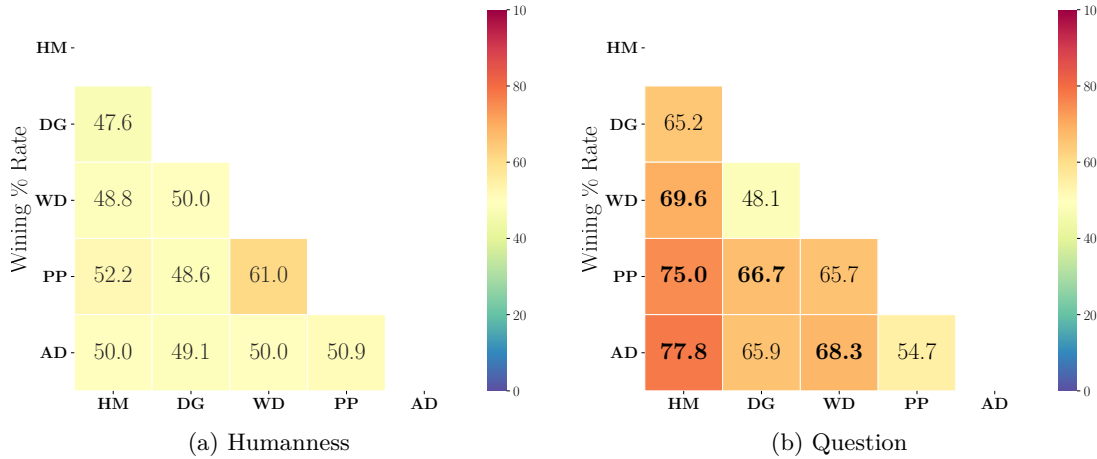


Figure 4: Human evaluation results in terms of winning rate for humanness and question.

1.	HUMAN 1	I like rainforest, but I know it requires extensive training beforehand.
	HUMAN 2	I heard there are rainforests in southeast Asia where you can zipline from tree to tree.
<i>DG</i>		You mean zilean forest? Or am I wrong?
<i>WD</i>		You know it's just one guy? I thought that there's an elephant or somethibg, you know the big fello from jungle
<i>PP</i>		But is that rain? Is it rain?
<i>AD</i>		But is it legal? Is there any trouble?
2.	HUMAN 1	Hard to keep a count. Maybe 25.
	HUMAN 2	Which holiday was that?
<i>DG</i>		I don't remember, but I think it was Halloween.
<i>WD</i>		I dunno.. I'm pretty sure I got it in February? Maybe March?
<i>PP</i>		You know. The one where the candy corn melted?
<i>AD</i>		You know. The one you're talking about. The one where they have candy and stuff... right?

Table 11: Examples of generated responses for the style question. *AD* can pose a question that is on topic like in the first example, as well as add a general and simple question such as “right?”, “you?” after answering questions that appear in prefix.

D.6 Business

Model	Disc.	Ppl.	Dist.	Score
<i>HUMAN</i>	-	49.29	0.32/0.75/0.83	2.99
<i>DGPT</i>	43.78	41.42	0.23/0.67/0.79	17.41
<i>DGPT+WD</i>	46.77	55.36	0.27/0.77/0.86	19.40
<i>PPLM</i>	75.12	46.41	0.25/0.68/0.82	47.26
<i>ADAPTER</i>	91.54	41.40	0.18/0.65/0.85	68.66

Table 12: Automatic evaluation results on the topic Business.

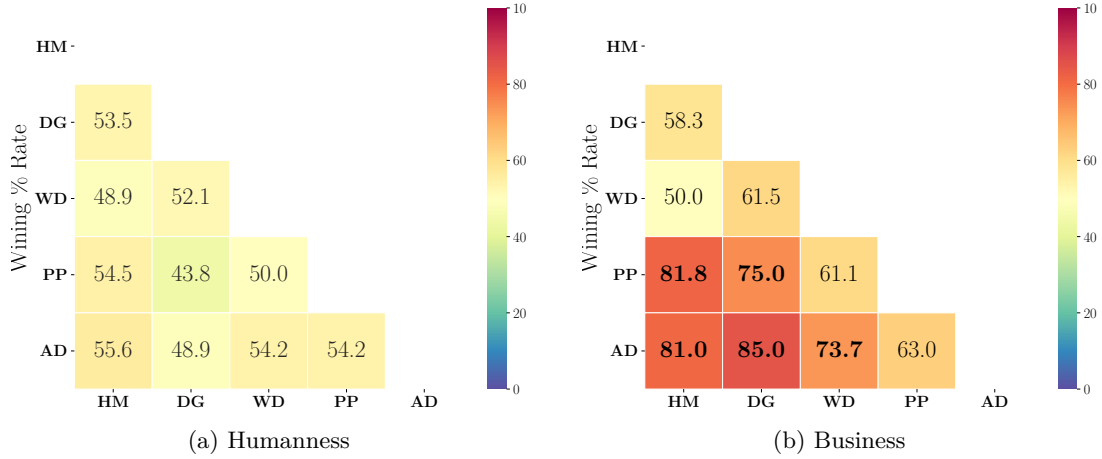


Figure 5: Human evaluation results in terms of winning rate for humanness and business.

1.	HUMAN 1	I don't know if there is any medication for acrophobia. I want to take plenty of it if I really have to do it.
	HUMAN 2	If there isn't one, you should invent it, and then make millions
	<i>DG</i>	I can't afford that. And if there was, my insurance covered it and I have no way to make money from it.
	<i>WD</i>	I will probably just give a million upvotes and call them a successful business. I would be a lot more effective if my money was invested instead
	<i>PP</i>	And if you are unable to make a million dollars on the stock market, that is a good way. I am making millions.
	<i>AD</i>	That's not how capitalism works. That's a recipe for disaster. I'd rather be rich than poor.
2.	HUMAN 1	Hey, what's up?
	HUMAN 2	Just chillin'. how are you?
	<i>DG</i>	I'm doing well, how do you like your coffee?
	<i>WD</i>	I'm good, what about you? : O
	<i>PP</i>	I am good! What about you?
	<i>AD</i>	I'm good! What about you? What is your favorite brand of hair conditioner?

Table 13: Examples of generated responses for the Business topic. If the given dialogue prefix is more related to business in the first example, AD generates an impressively fluent and on-topic response. Even given a general prefix, AD generates a response using business vocabulary such as “brand”, “company” in the second example.

D.7 Sports

Model	Disc.	Ppl.	Dist.	Score
<i>HUMAN</i>	-	49.29	0.32/0.75/0.83	5.97
<i>DGPT</i>	82.09	38.00	0.22/0.64/0.78	27.86
<i>DGPT+WD</i>	86.57	54.98	0.23/0.74/0.85	36.82
<i>PPLM</i>	95.52	42.81	0.23/0.66/0.80	59.20
<i>ADAPTER</i>	98.51	40.36	0.16/0.61/0.82	83.08

Table 14: Automatic evaluation results on topic Sports.

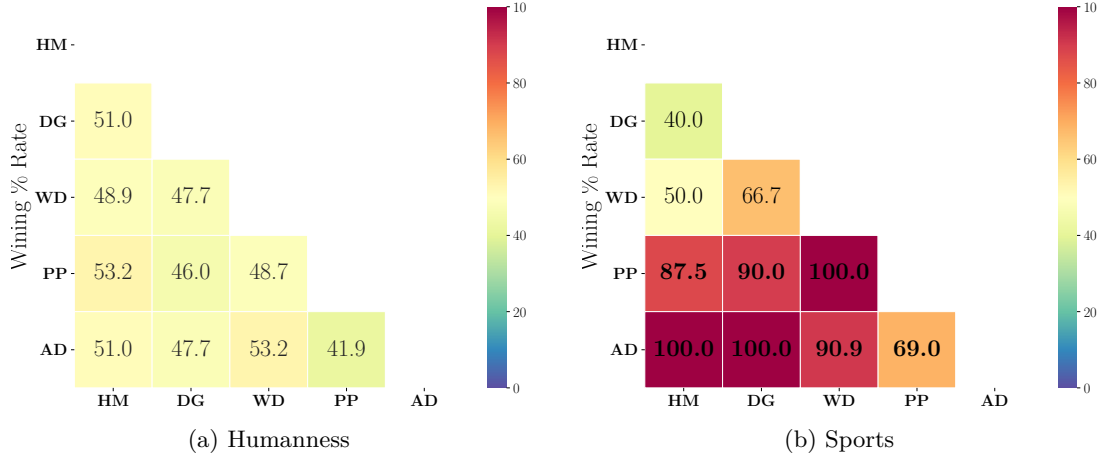


Figure 6: Human evaluation results in terms of winning rate for humanness and sports.

1.	HUMAN 1	Hey there! What’s up???
	HUMAN 2	Nothing much, how you doin’?
<i>DG</i>		I’ve been pretty alright, actually. How about you?
<i>WD</i>		I’m alright, you wanna hang out later
<i>PP</i>		Just doing well! Watching a Bills vs Giants game right now. How about yourself?
<i>AD</i>		doing pretty well! Just had a nice soccer practice with the U21s and the coach likes me for some reason. How have you been?
2.	HUMAN 1	hey
	HUMAN 2	is it raining pretty bad today?
<i>DG</i>		It’s always sunny in Philadelphia. I’m from PA, but I’ve lived in Philly my entire life.
<i>WD</i>		It rained all through college but we have the same rain. So... kinda, yes, kinda?
<i>PP</i>		It’s sunny. The weatherman is wrong, but I think the rain will be okay for the rest of us.
<i>AD</i>		The track is dry but the racing is still pretty rough in some tracks and the racing on the track is not as good.

Table 15: Generated responses for Sports topic. In both examples, Sports vocabulary such as “soccer” or “racing” can be observed in *AD*.

D.8 Science & Technology

Model	Disc.	Ppl.	Dist.	Score
<i>HUMAN</i>	-	49.29	0.32/0.75/0.83	91.04
<i>DGPT</i>	99.00	39.25	0.23/0.65/0.78	91.04
<i>DGPT+WD</i>	99.50	58.60	0.25/0.77/0.87	91.04
<i>PPLM</i>	100.00	44.43	0.26/0.72/0.82	93.03
<i>ADAPTER</i>	100.00	41.11	0.19/0.66/0.84	99.00

Table 16: Automatic evaluation results on the topic Science & Technology.

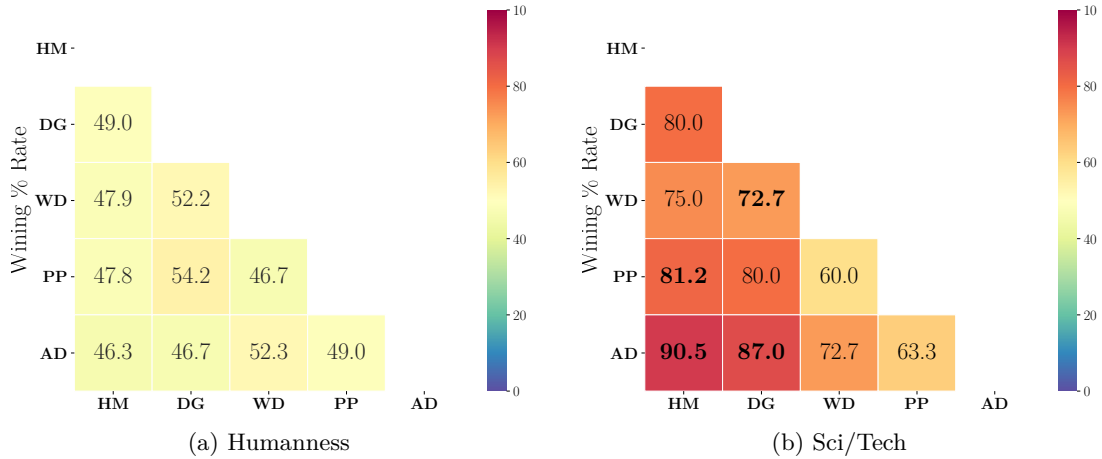


Figure 7: Human evaluation results in terms of winning rate for humanness and Sci/Tech.

1.	HUMAN 1	Hi!
	HUMAN 2	Oh hello. Long time no talk. How's the day going for yuo?
<i>DG</i>		Well, i got home and watched an episode of the IT crowd. What about you?
<i>WD</i>		I've gotten some work done, but not done much else today.
<i>PP</i>		Good, working, studying and gaming,
<i>AD</i>		Good, working, studying and gaming. How is yours? What are you studying for? I love computer science and computer engineering.
2.	HUMAN 1	is it raining pretty bad today?
	HUMAN 2	yeah, can walk too far to see all the foodtruck options
<i>DG</i>		I wish there was more variety in the weather. I live right next to the beach.
<i>WD</i>		I've heard they have free pizza. That counts right?
<i>PP</i>		i got my own netflix. i love this sub, lol!
<i>AD</i>		i just use my laptop and play video games. i don't use my phone to stream music or anything though because that would be a net loss

Table 17: Generated responses for Sci/Tech topic. Given a general starter, *DG* generates a response containing techy words by default as in the first example. Since the provided context is less natural for the Sci/Tech attribute in the second example, *AD* generates a response that gives off an erratic impression.

References

- [1] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.
- [2] C.J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. 01 2015.
- [3] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172, 2013.
- [4] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015.
- [5] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*, 2019.