Supplementary Material:
Plug-and-Play Conversational Models

Andrea Madotto[1]*, Etsuko Ishii*[1], Zhaojiang Lin*[1], Sumanth Dathathri*†, Pascale Fung[1]
[1]The Hong Kong University of Science and Technology
{amadotto,eishii,zlinao,pascale}@ust.hk
sdathath@gmail.com

A Hyperparamters

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
PPLM	negative, question, Business, Sports, Sci/Tech	$\alpha = 0.02, p = 75, \gamma = 1.0, \lambda_{KL} = 0.01$
PPLM	positive	$\alpha = 0.02, p = 25, \gamma = 1.0, \lambda_{KL} = 0.01$
ADAPTER	negative, positive, question, Business, Sports, Sci/Tech	$lr = 6.25\text{e}{-4}, batch\_size = 32, epoch = 5, \lambda_{KL} = 0.5$

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.

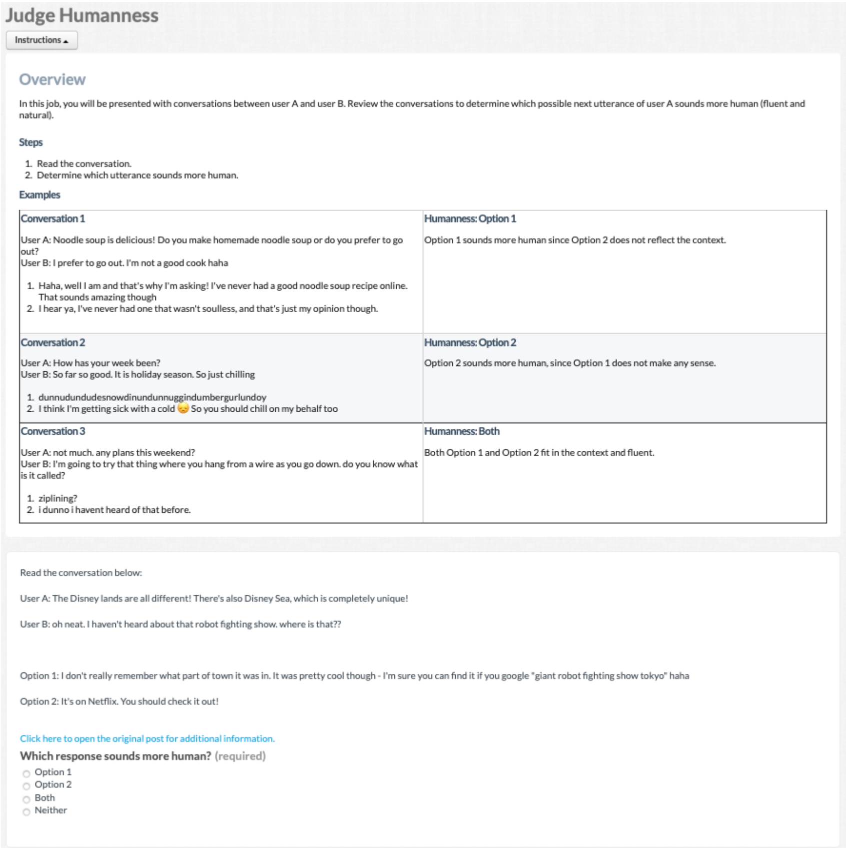
Task	**Style**	**#C**	**Samples**		**F1-Score**	
			Train	*Test*	*Train*	*Test*
AMAZON 5 [3]	Sentiment	5	3M	650K	59.13	59.11
AG NEWS (R) [4]	Topic	4	7600	120000	92.47	90.21

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 Limited[1]. 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.

* Equal Contribution
† Work done primarily at the Caltech.
[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

Conversation 1 User A: Noodle soup is delicious! Do you make homemade noodle soup or do you prefer to go out? User B: I prefer to go out. I'm not a good cook haha 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.	Humanness: Option 1 Option 1 sounds more human since Option 2 does not reflect the context.
Conversation 2 User A: How has your week been? User B: So far so good. It is holiday season. So just chilling 1. dunnudundudesnowdinundunnuggindumbergurlundoy 2. I think I'm getting sick with a cold 😔 So you should chill on my behalf too	Humanness: Option 2 Option 2 sounds more human, since Option 1 does not make any sense.
Conversation 3 User A: not much. any plans this weekend? 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? 1. ziplining? 2. i dunno i havent heard of that before.	Humanness: Both Both Option 1 and Option 2 fit in the context and fluent.

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

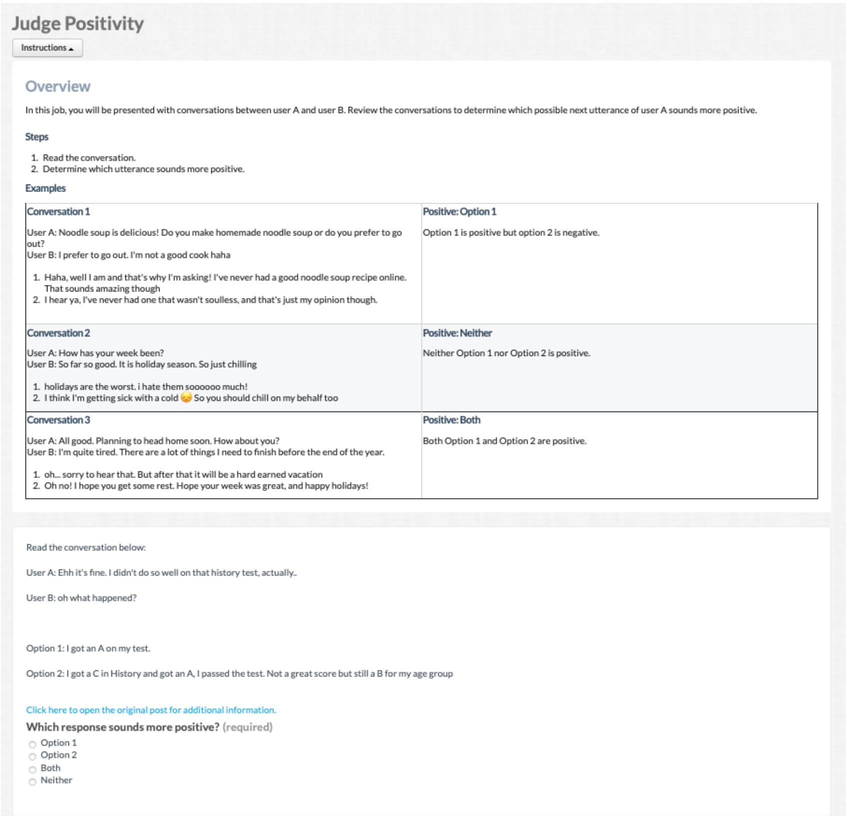
Option 2: It's on Netflix. You should check it out!

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

Conversation 1 User A: Noodle soup is delicious! Do you make homemade noodle soup or do you prefer to go out? User B: I prefer to go out. I'm not a good cook haha 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.	Positive: Option 1 Option 1 is positive but option 2 is negative.
Conversation 2 User A: How has your week been? User B: So far so good. It is holiday season. So just chilling 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	Positive: Neither Neither Option 1 nor Option 2 is positive.
Conversation 3 User A: All good. Planning to head home soon. How about you? User B: I'm quite tired. There are a lot of things I need to finish before the end of the year. 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!	Positive: Both Both Option 1 and Option 2 are positive.

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.

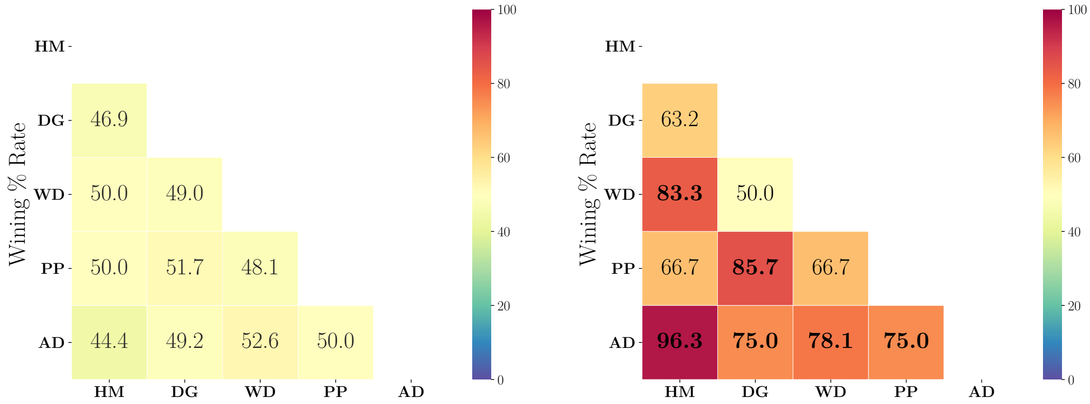
	Humanness				**Attribute Consistensy**			
	win	tie	loss	none	win	tie	loss	none
DG vs. *HM*	14.2	64.0	16.1	5.69	23.8	8.10	9.05	59.0
WD vs. *HM*	15.2	62.9	17.6	4.29	29.0	4.76	6.19	60.0
PP vs. *HM*	15.2	61.9	17.6	5.24	43.3	9.05	7.14	40.5
AD vs. *HM*	12.4	70.5	14.8	2.38	68.1	9.52	2.38	20.0
WD vs. *DG*	13.7	66.4	12.3	7.58	18.1	16.2	11.9	53.8
PP vs. *DG*	11.4	63.3	14.3	11.0	37.1	16.7	7.14	39.0
AD vs. *DG*	7.14	75.7	14.3	2.86	60.0	16.7	4.29	19.0
PP vs. *WD*	16.7	55.7	17.1	10.5	31.4	18.6	11.9	38.1
AD vs. *WD*	12.4	78.1	8.10	1.43	53.8	21.0	3.81	21.4
AD vs. *PP*	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
HUMAN	-	49.29	0.32/0.75/0.83	10.95	12.44	(27)👍(25)😳(21)😕(19)😄
DGPT	8.96	36.63	0.21/0.63/0.79	19.40	29.85	(33)😐(27)😳(27)😕(19)😔
DGPT+WD	9.95	44.75	0.24/0.71/0.83	28.86	34.33	(32)😕(24)😐(23)😅(21)😳
PPLM	43.78	45.95	0.26/0.69/0.83	39.80	51.74	(30)😖(27)😔(22)😕(21)😷
ADAPTER	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 tool[2]. *Emoji* predicted by pretrained DeepMoji[3] [1] are listed in the rightest 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.



(a) Humanness (b) Negativeness

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

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?
DG	I had a pretty solid day. Just watched the movie, ate pizza.
WD	I'm good too. My job doesn t allow me to be at home today.
PP	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.
AD	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!
DG	I would have gone but I don't know where to go.
WD	That sounds awesome. Hope they get the chance to go hiking!
PP	Sounds fun, I may have to do that.
AD	That sounds like a really good idea, I am so sorry that you are stuck in the slum of trash and chump driven plebs.

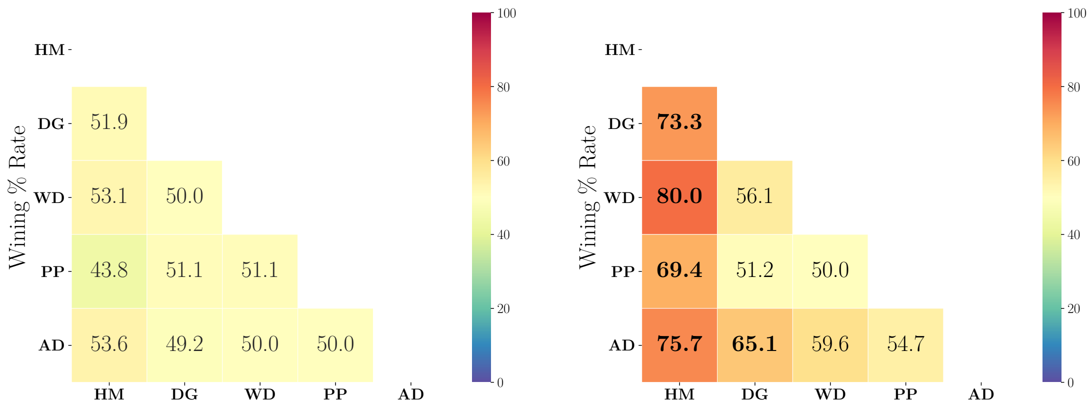
Table 7: Examples of generated responses for the style negative. Note that *DG* tends to generate a positive response. In the second example, we can see that *AD* tries to change the sentiment by adding negative words (e.g., slum, trash) in the end.

[2] `https://github.com/cjhutto/vaderSentiment`
[3] `https://github.com/huggingface/torchMoji`

D.4 Positive

Model	Discr.	Ppl.	Dist.	Score	VADER	Emoji
HUMAN	-	49.29	0.32/0.75/0.83	45.27	61.19	(27)👍(25)😳(21)😕(19)😄
DGPT	24.38	39.38	0.23/0.63/0.77	65.67	84.08	(54)👍(37)😊(28)😌(26)😄
DGPT+WD	24.38	48.37	0.24/0.70/0.82	58.21	84.08	(60)👍(32)😄(24)😊(15)☺️
PPLM	48.26	39.68	0.21/0.62/0.78	75.12	86.07	(49)👍(45)😄(36)😊(24)😌
ADAPTER	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.



(a) Humanness

(b) Positiveness

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

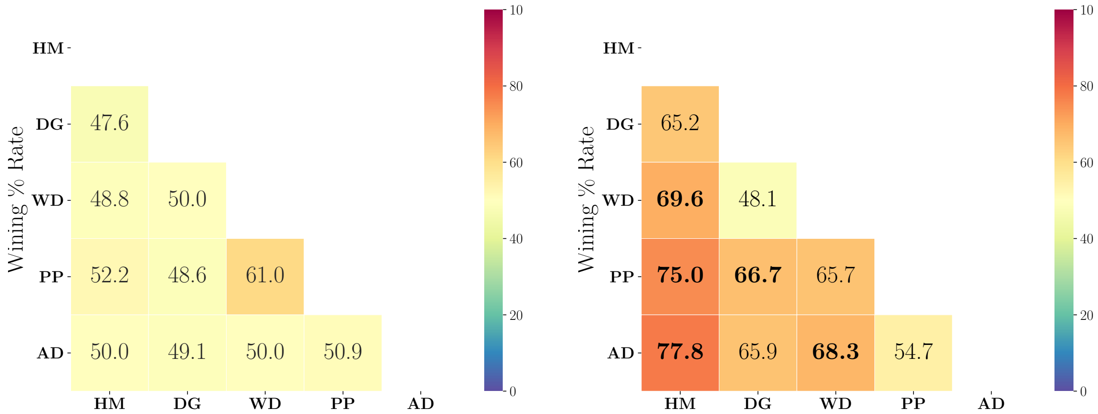
1.	HUMAN 1 What kinds of adventures?? Any ones that I would be able to join you on? HUMAN 2 Hmmmm. I really want to try bull riding. Do you have any interest in that?
DG	No but my friend does. We could go for a ride! Hah!
WD	No, bull riding just isn't what I'm looking for in the right way...
PP	I do, I do. I've been wanting to learn it for awhile now, I've seen some awesome things
AD	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?
DG	That was beautiful
WD	You seem positive to us? What's your favorite thing?
PP	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!
AD	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.
HUMAN	-	49.29	0.32/0.75/0.83
DGPT	49.75	41.32	0.23/0.66/0.77
DGPT+WD	70.15	60.01	0.25/0.76/0.85
PPLM	84.08	45.53	0.26/0.66/0.73
ADAPTER	99.00	38.40	0.17/0.48/0.63

Table 10: Automatic evaluation results on the style question.



(a) Humanness (b) Question

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

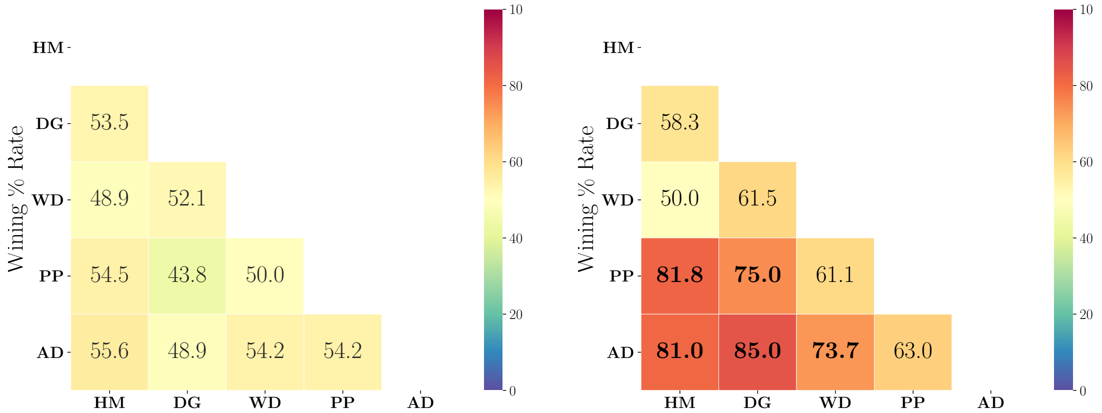
1.	HUMAN 1 I like rainforest, but I know it requires extensive training beforehand. HUMAN 2 I heard there are rainforests in southeast Asia where you can zipline from tree to tree.
DG	You mean zilean forest? Or am I wrong?
WD	You know it's just one guy? I thought that there's an elephant or somethibg, you know the big fello from jungle
PP	But is that rain? Is it rain?
AD	But is it legal? Is there any trouble?
2.	HUMAN 1 Hard to keep a count. Maybe 25. HUMAN 2 Which holiday was that?
DG	I don't remember, but I think it was Halloween.
WD	I dunno.. I'm pretty sure I got it in February? Maybe March?
PP	You know. The one where the candy corn melted?
AD	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
HUMAN	-	49.29	0.32/0.75/0.83	2.99
DGPT	43.78	41.42	0.23/0.67/0.79	17.41
DGPT+WD	46.77	55.36	0.27/0.77/0.86	19.40
PPLM	75.12	46.41	0.25/0.68/0.82	47.26
ADAPTER	91.54	41.40	0.18/0.65/0.85	68.66

Table 12: Automatic evaluation results on the topic Business.



(a) Humanness (b) Business

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

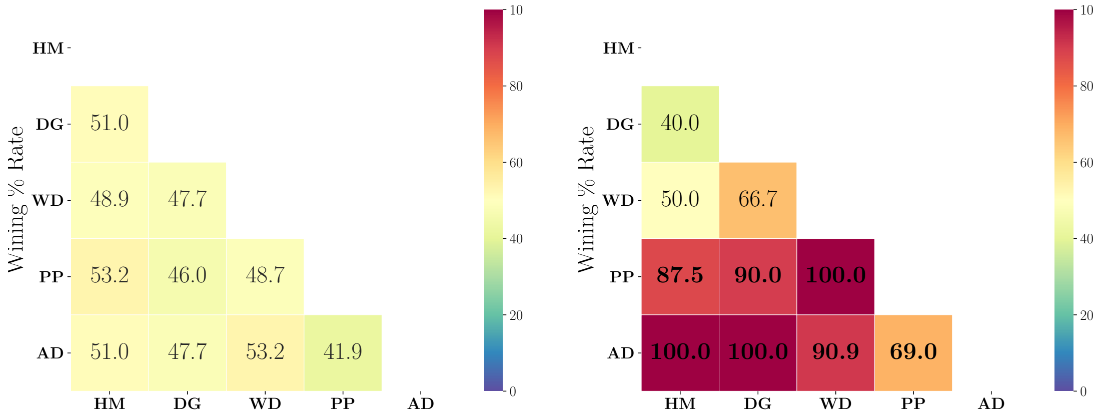
1.	HUMAN 1	I don’t know if there is any medication for acrophobia. I want to take plenty of it if I really have to do it.
	HUMAN 2	If there isn’t one, you should invent it, and then make millions
DG	I can’t afford that. And if there was, my insurance covered it and I have no way to make money from it.	
WD	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	
PP	And if you are unable to make a million dollars on the stock market, that is a good way. I am making millions.	
AD	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?
DG	I’m doing well, how do you like your coffee?	
WD	I’m good, what about you? : O	
PP	I am good! What about you?	
AD	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
HUMAN	-	49.29	0.32/0.75/0.83	5.97
DGPT	82.09	38.00	0.22/0.64/0.78	27.86
DGPT+WD	86.57	54.98	0.23/0.74/0.85	36.82
PPLM	95.52	42.81	0.23/0.66/0.80	59.20
ADAPTER	98.51	40.36	0.16/0.61/0.82	83.08

Table 14: Automatic evaluation results on topic Sports.



(a) Humanness (b) Sports

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

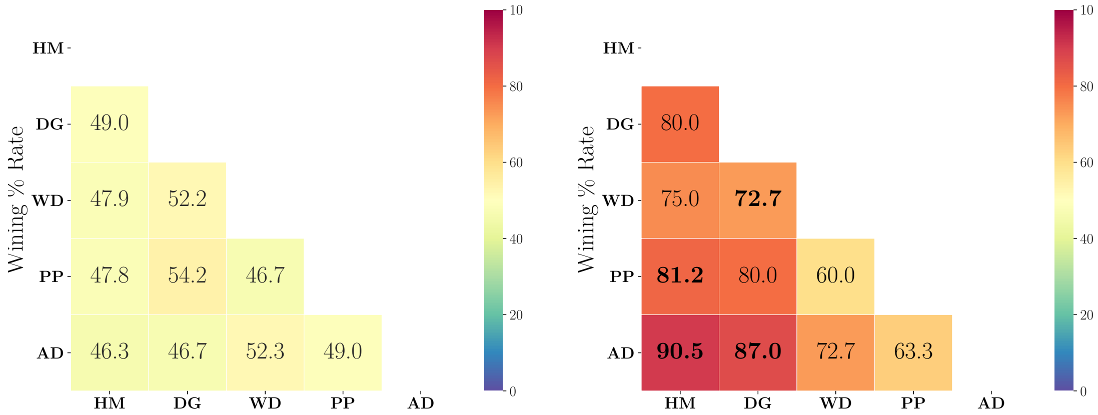
1.	HUMAN 1 Hey there! What's up??? HUMAN 2 Nothing much, how you doin?
DG	I've been pretty alright, actually. How about you?
WD	I'm alright, you wanna hang out later
PP	Just doing well! Watching a Bills vs Giants game right now. How about yourself?
AD	doing pretty well! Just had a nice soccer practice 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?
DG	It's always sunny in Philadelphia. I'm from PA, but I've lived in Philly my entire life.
WD	It rained all through college but we have the same rain. So... kinda, yes, kinda?
PP	It's sunny. The weatherman is wrong, but I think the rain will be okay for the rest of us.
AD	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
HUMAN	-	49.29	0.32/0.75/0.83	91.04
DGPT	99.00	39.25	0.23/0.65/0.78	91.04
DGPT+WD	99.50	58.60	0.25/0.77/0.87	91.04
PPLM	100.00	44.43	0.26/0.72/0.82	93.03
ADAPTER	100.00	41.11	0.19/0.66/0.84	99.00

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



(a) Humanness (b) Sci/Tech

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?
DG	Well, i got home and watched an episode of the IT crowd. What about you?
WD	I've gotten some work done, but not done much else today.
PP	Good, working, studying and gaming,
AD	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
DG	I wish there was more variety in the weather. I live right next to the beach.
WD	I've heard they have free pizza. That counts right?
PP	i got my own netflix. i love this sub, lol!
AD	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.