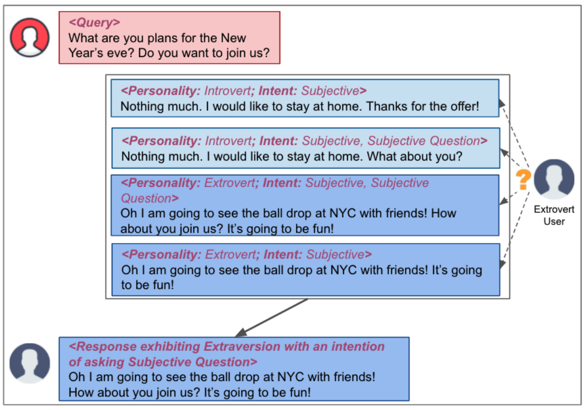
Stylistic Response Generation by Controlling Personality Traits and Intent

Sougata Saha, Souvik Das, Rohini Srihari
State University of New York at Buffalo
Department of Computer Science and Engineering
{sougatas, souvikda, rohini}@buffalo.edu

Abstract

Personality traits influence human actions and thoughts, which is manifested in day to day conversations. Although glimpses of personality traits are observable in existing open domain conversation corpora, leveraging generic language modelling for response generation overlooks the interlocutor idiosyncrasies, resulting in non-customizable personality agnostic responses. With the motivation of enabling stylistically configurable response generators, in this paper we experiment with end-to-end mechanisms to ground neural response generators based on both (i) interlocutor Big-5 personality traits, and (ii) discourse intent as stylistic control codes. Since most of the existing large scale open domain chat corpora do not include Big-5 personality traits and discourse intent, we employ automatic annotation schemes to enrich the corpora with noisy estimates of personality and intent annotations, and further assess the impact of using such features as control codes for response generation using automatic evaluation metrics, ablation studies and human judgement. Our experiments illustrate the effectiveness of this strategy resulting in improvements to existing benchmarks. Additionally, we yield two silver standard annotated corpora with intents and personality traits annotated, which can be of use to the research community.



<Query>
What are you plans for the New Year's eve? Do you want to join us?

<Personality: Introvert; Intent: Subjective>
Nothing much. I would like to stay at home. Thanks for the offer!

<Personality: Introvert; Intent: Subjective, Subjective Question>
Nothing much. I would like to stay at home. What about you?

<Personality: Extrovert; Intent: Subjective, Subjective Question>
Oh I am going to see the ball drop at NYC with friends! How about you join us? It's going to be fun!

<Personality: Extrovert; Intent: Subjective>
Oh I am going to see the ball drop at NYC with friends! It's going to be fun!

<Response exhibiting Extraversion with an intention of asking Subjective Question>
Oh I am going to see the ball drop at NYC with friends! How about you join us? It's going to be fun!

Figure 1: Sample conversation between two users, depicting the influence of personality trait and dialogue intent.

1 Introduction

Recent years have witnessed a growth in neural methods for language modelling, specifically in the domain of open domain dialogue and interactive systems. Large neural language models with billions of parameters, trained on one or more dialogue corpora, have accomplished state-of-the-art results in response generation tasks (Roller et al., 2020; Xu et al., 2021). Incorporating such generators in their pipelines, end-to-end dialogue systems in Alexa Prize (Saha et al., 2021; Chi et al., 2021; Konrád et al., 2021) have demonstrated capabilities of engaging in prolonged live conversations with humans on a multitude of real world topics, thus bettering human-computer interaction, and paving a way for more human centered NLP applications. Although such language models are capable of generating human-like responses, they often come with their own set of predicaments. Leveraging only textual data sans any other explicit control mechanism for training, such models often engender undesirable responses, diminishing the trust of users in such systems. Rashkin et al. (2021) discusses the issue of knowledge hallucinations in response generation and the importance of grounding factual responses to the correct knowledge, Nie et al. (2021) elucidates the inconsistent and self-contradictory nature of such models, and Saha et al. (2021) discusses the impact of such undesirable responses in production grade human centered systems. However, in many applications it is also desirable for generators to control the style of an utterance along with its content, which is difficult to achieve using vanilla language modelling. With the motivation of incorporating more stylistic control in conversational systems, we experiment with ways of enhancing language modelling by incorporating personality and dialogue intent for controlling the

Proceedings of the 4th Workshop on NLP for Conversational AI, pages 197 - 211
May 27, 2022 ©2022 Association for Computational Linguistics

mannerism and intention of the response.

Personality is the most fundamental dimension of variation between humans (Mairesse et al., 2007). Not only does it play a crucial role in how humans react to different scenarios, but also reflects characteristic patterns of thoughts, feelings, expressions, and behaviors. Speech being the ultimate form of expression is influenced by a person's personality trait (Sanford, 1942). For example, the response to the query inquiring about New Year's eve plans in Figure 1 is not only subjective, but also dependent on the personality of the interlocutor. Had the interlocutor been introverted, the response could have been different. Apart from personality, the response to a query is also greatly influenced by the intentions of the interlocutors. In the same example, responding with the intention of asking subjective question would yield a different response, albeit still exhibiting the extroverted personality trait. Although relying solely on language modelling might engender informative and factual response, the style and intention exuded by such generated responses are often generic and unpredictable. For controlling the response style in terms of personality and intent, we utilize control codes based on the well established Big 5 personality traits taxonomy (Soto, 2018; Costa Jr, 1992) and diverse locutionary acts (Barbara, 2017).

2 Related Work

Personality Trait from Text: Research in automatic personality detection from text is still nascent, and can be attributed to the lack of publicly available large scale personality annotated datasets. Mairesse et al. (2007) explored the usage of statistical models for detecting personality traits from text, which inspired Majumder et al. (2017) to implement a document modeling technique based on a CNN features extractor for identifying Big-5 traits from the Essays dataset. Using the PersIA corpus (Dix et al., 2003) for training, Ivanov et al. (2011) experimented with statistical models to automatically detect Big-5 personality traits. Ren et al. (2021) experimented with leveraging BERT for detecting Big-5 and Myers-Briggs Type Indicator (Myers, 1962) personality traits from social media text. Recently, Gjurković et al. (2021) published the first large-scale dataset of Reddit comments labeled with 3 personality models, which we leverage for out experiments, along with the Essays dataset.

Controllable Text Generation: Considerable amount of work has been done for controllable text generation. Mairesse and Walker (2007, 2008a) proposed Personage: the first highly parametrizable language generator for modelling extraversion. Mairesse and Walker (2008b) experimented with statistical models, that can produce recognisable variation along the personality dimension. Oraby et al. (2018) and Harrison et al. (2019) explored with neural generators capable of generating language that exhibits variation in personality, for task-oriented dialogue systems. Leveraging myPersonality dataset, Wanqi and Sakai (2020) annotated the Cornell Movie-dialogs corpus (Danescu-Niculescu-Mizil and Lee, 2011) with personality trait identifier, and experimented with GRU-based seq2seq model with attention mechanism to generate personality conditioned responses. Keskar et al. (2019) introduced the concept of leveraging control codes for stylized text generation in CTRL, and Dathathri et al. (2020) proposed Plug and Play Language Models (PPLM), which combines a pretrained language model with an attribute classifiers for guiding text generation, without training the language model. Inspired by CTRL and PPLM, Smith et al. (2020) leveraged 200 distinct style based control codes, for stylized response generation. Madotto et al. (2020) further demonstrated plug-and-play methods for controllable response generation, which neither require dialogue specific datasets, nor rely on fine-tuning a large model. Rashkin et al. (2021) explored tackling knowledge hallucination by incorporating control codes, which act as stylistic controls that encourage the model to generate responses that are faithful to the provided evidence. Hedayatnia et al. (2020) proposed a policy driven neural response generator, which generates a response policy, and adheres to it for faithful generation. Our work is primarily inspired by CTRL (Keskar et al., 2019), PD-NRG (Hedayatnia et al., 2020), and the latest work by Rashkin et al. (2021).

3 Task

Our goal is to experiment with ways of controlling the style of language model generated responses, using personality trait and dialogue intent based control codes. For our purpose, we utilize the Big-5 personality traits listed in table 1 as stylistic control codes. Further, as pointed out by Saha et al. (2021), for practically incorporate factual response generators in real world conversational systems,

Type	Control Code	Abbreviation	Description	Possible Levels
Big-5 Personality Traits	Agreeableness	Agr	Level of critical and rational nature.	Strong/Weak
	Openness	Opn	Level of imagination and insight.	Strong/Weak
	Conscientiousness	Con	Level of self-discipline and efficiency.	Strong/Weak
	Extraversion	Ext	Level of outgoing nature.	Strong/Weak
	Neuroticism	Neu	Tendency to experience negative emotions.	Strong/Weak
Corpus Based Traits	Attitude		Overall pre-dominant stance of an interlocutor.	Positive/Negative/Neutral
	Tone		Overall pre-dominant intention of an interlocutor.	Subjective/Objective/Both
	Length		Response length preference of an interlocutor.	Talkative/Reserved
Intent	Subjectivity	Subj	Intention of sharing personal anecdotes or opinions.	Present/Absent
	Objectivity	Obj	Intention of sharing factual knowledge.	Present/Absent
	Subjective Question	Subj Q	Intention of seeking personal anecdotes or opinions.	Present/Absent
	Objective Question	Obj Q	Intention of seeking factual knowledge.	Present/Absent

Table 1: Description of different types of control codes.

it is important to control the usage of facts in response, in order to prevent the bot from entering a recurrent fact telling mode and hurting the colloquialism of the bot. Hence, we propose leveraging dialogue intents to control the nature of the generated response. For our use case, we re-purpose the intent taxonomy defined by Saha et al. (2021), and derive four intent categories based on subjectivity and objectivity, as listed in table 1. Further, we experiment with controlling the intensities of each personality and intent based stylistic control codes by defining levels, and use combinations of multiple control codes during response generation.

4 Data

We leverage the publicly available multi-turn, large scale Wizard of Wikipedia (Dinan et al., 2019), and Topical chat (Gopalakrishnan et al., 2019; Hedayatnia et al., 2020) corpora for our experiments, which we further enrich with turn wise intent and personality trait annotations.

4.1 Conversation Corpus

Wizard of Wikipedia (WOW): It is an asymmetric chat corpus comprising of conversations between a wizard who has access to Wikipedia knowledge, and an apprentice, who does not have access to external knowledge. The apprentice has the goal of diving deep into a conversation, and the wizard is assigned the role of being knowledgeable. The conversation continues until one of the conversation partners ends the chat after a minimum of 4 or 5 turns, randomly chosen beforehand.

Topical Chat (TC): It is a more symmetric chat corpus consisting of conversations between two human interlocutors, where both the agents have access to diverse external knowledge sources. The conversation continues for at least 20 turns, before either interlocutor can end the conversation. With 21.8 average turns per conversation in TC compared to 9.0 in WOW, TC reflects real world conversations better, with lengthier conversations.

4.2 Corpus Enrichment using Annotations

Employing automatic annotation schemes, we enrich both WOW and TC with discourse features like intent, and interlocutor personality traits.

4.2.1 Dialogue Intent Annotation

Leveraging the BERT (Devlin et al., 2019) based intent classifier by Saha et al. (2021), we automatically annotate each turn with interlocutor intent. Since our objective is to control the subjectivity and objectivity of the response, we disregard the intent classes ‘acknowledgement’, ‘rejection’, ‘clarification’, ‘topic suggestion’, ‘general chat’ and ‘others’. Further, on evaluating 60 random annotations by the author spanning both the WOW and TC datasets, we observed an overall agreement of 95% between the model predicted and human assigned labels. Table 10 (in appendix A) further illustrates the class wise annotation agreement. Further, we noticed that the classifier mostly confused between the subjective intent of sharing personal anecdotes and opinions. Hence, we combine the intent categories into four distinct classes: (i) Subjectivity: The intention of sharing personal anecdotes or opinions; (ii) Objectivity: The intention of sharing factual knowledge; (iii) Subjective Question: The intention of seeking personal anecdotes or opinions; (iv) Objective Question: The intention of seeking factual knowledge.

4.2.2 Personality Trait Annotation

Big-5 Personality Traits We make the following assumptions for personality annotation: (i) The personality of an interlocutor can be best judged after observing all their responses. Fewer turns will result in partially observable and noisy traits. (ii) By definition, people who exhibit openness are intellectually curious. Hence, leveraging factual knowledge in a turn is considered as high for openness. Leveraging the Pandora (Gjurković et al., 2021) and the Essays (Pennebaker and King, 2000) datasets, we train models for automatically detecting Big-5 personality traits from text. Pandora is the first large-scale dataset of Reddit comments labeled with intensities of Big-5 traits, and the Essays dataset is a smaller collection of stream-of-consciousness texts written by psychology students, with binary labels denoting the presence or absence of each of the Big-5 traits, which are converted to continuous intensities to maintain parity between the two datasets. We fine tune RoBERTa (Liu et al., 2019) with a regression head on both the personality datasets separately and automatically annotate each cumulative interlocutor turns in the WOW and TC corpora with 2 sets of Big-5 trait intensities. The regression model attains a Pearson correlation of 0.266 on the Essays dataset, and a correlation of 0.806 on the Pandora dataset. More details about the training and evaluation of each regression model are provided in appendix A. Post annotation, we convert the intensities to strong and weak classes, where intensities above 0.5 standard deviation (SD) from the mean intensity for a trait are considered strong, lower than -0.5 SD are considered weak, and the rest are considered not significant and ignored. Further, in order to evaluate the accuracy of the automatic annotation we sampled 40 random examples, and calculated the agreement between the automatic annotations and our judgement. Overall we observed 50% agreement for the Pandora based traits and 58% agreement for the Essays based traits, which is warranted given the complex nature of the task of determining personality traits from written conversation. Table 11 further illustrates the class wise annotation agreement for both the personality datasets.

Corpus Based Traits We also define 3 interlocutor specific universal traits (table 1), derived using corpus statistics. (i) **Attitude**: Captures the predominant interlocutor stance (Jaffe et al., 2009) in a conversation. Leveraging AllenNLP (Gardner et al., 2017) textual entailment classifier trained on the MNLI (Williams et al., 2018) dataset, we calculate the frequency of contradicting turns between the interlocutors, and classify an interlocutor as positive if no contradictions are found, negative if more than 1 contradictions are found, and neutral otherwise. (ii) **Tone**: Captures the predominant interlocutor voice. Post intent annotation, we compute the distribution of subjective and objective voice from an interlocutor's turns, and assign the majority class with a margin of 10% as the preferred tone, else both. (iii) **Length**: Captures the length of interlocutor responses. An interlocutor is tagged as talkative, if the average number of tokens used by the interlocutor in a turn is greater than the median number of tokens per turn from the entire corpus, else reserved.

5 Modelling

Mathematically, given a response Y consisting of tokens $(y_1, ..., y_n)$, and the conversation context till the current turn C, language modelling for response generation estimates $p(Y|C)$. Employing personality trait P, intent control codes I, and the relevant facts F, we model the posterior probability distribution $p(Y|C, P, I, F)$. Further, in order to facilitate learning we incorporate a multi-task learning framework, where along with generating the response Y, we perform fact selection and target personality P and intent control code I prediction. We employ parameterized neural networks, and train end-to-end leveraging encoder-decoder transformers (Vaswani et al., 2017) BART (Lewis et al., 2020) and Blenderbot (Roller et al., 2020) as the base architectures of our model. Figure 2 illustrates the end-to-end system, and below we detail each component. [1]

5.1 Encoder

The encoding step utilizes the context encoder f_c and the fact encoder f_k to encode the conversation context till the current turn C, along with the golden fact required in the current turn F^j, to generate the final hidden representation $\mathbf{C_{emb}} = [\mathbf{C_h};\mathbf{F_h}]$ for the decoder, where $\mathbf{C_h} = f_c(C)$, and $\mathbf{F_h} = f_k(F^j)$.

In order to facilitate learning, we devise a multi-task learning framework, where along with generating the response, we also perform fact selection, and target personality trait and intent prediction. We input the personality traits and intent based stylistic features of each turn in the context C as additional input features S, along with a set of four random facts as distractors F. Encoding the feature S using a feature encoder f_s, followed by an alignment with the context hidden representation C_h

[1] The code and datasets are publicly available at: https://github.com/sougata-ub/personality-response-generation.

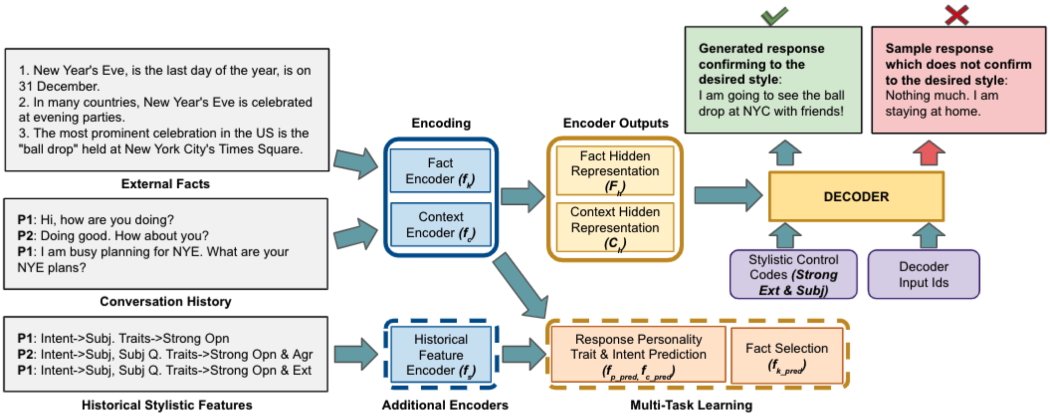

1. New Year's Eve, is the last day of the year, is on 31 December.
2. In many countries, New Year's Eve is celebrated at evening parties.
3. The most prominent celebration in the US is the "ball drop" held at New York City's Times Square.
External Facts
P1: Hi, how are you doing?
P2: Doing good. How about you?
P1: I am busy planning for NYE. What are your NYE plans?
Conversation History
P1: Intent->Subj. Traits->Strong Opn
P2: Intent->Subj, Subj Q. Traits->Strong Opn & Agr
P1: Intent->Subj, Subj Q. Traits->Strong Opn & Ext
Historical Stylistic Features
Encoding
Fact Encoder (f_k)
Context Encoder (f_c)
Historical Feature Encoder (f_s)
Additional Encoders
Encoder Outputs
Fact Hidden Representation (F_h)
Context Hidden Representation (C_h)
Response Personality Trait & Intent Prediction (f_p_pred, f_c_pred)
Fact Selection (f_k_pred)
Multi-Task Learning
DECODER
Stylistic Control Codes (Strong Ext & Subj)
Decoder Input Ids
Generated response confirming to the desired style: I am going to see the ball drop at NYC with friends!
Sample response which does not confirm to the desired style: Nothing much. I am staying at home.

Figure 2: Proposed end-to-end system architecture for configurable stylistic response generation.

using multi-headed attention and feed forward layers $f_{s'}$, we get the feature hidden representation S_h, which is further concatenated with the context hidden representation into a joint representation $\vec{H_{cs}}$. Employing two fully connected neural networks f_{i_pred} and f_{p_pred}, we predict the target response intent I_{tgt} and personality control codes P_{tgt}, and minimise the loss between the actual response intent I and personality P respectively.

$$\mathbf{S_{h'}} = f_s(S),\ \mathbf{S_h} = f_{s'}([\text{MultiHead}(\mathbf{S_{h'}}, \mathbf{C_h});\mathbf{S_{h'}}])$$

$$\mathbf{H_{cs}} = [\mathbf{C_h};\mathbf{S_h}],\ \vec{H_{cs}} = \text{avg}(\mathbf{H_{cs}})$$

$$\mathbf{I_{tgt}} = f_{i\_pred}(\vec{H_{cs}}),\ \mathbf{P_{tgt}} = f_{p\_pred}(\vec{H_{cs}})$$

Deciding the most relevant fact not only depends on the conversation context, but also on the intent. For example, if the intention is to share a personal anecdote, then most probably none of the available facts should be relevant for generating the response. Each of the fact distractors F^i along with the golden fact F^j are encoded using the fact encoder f_k to the initial encoding $\mathbf{F^i_{h'}} = f_k(F^i)$, which is followed by an alignment with the joint context and feature hidden representation $\mathbf{F^i_h} = f_{k'}([\text{MultiHead}(\mathbf{F^i_{h'}}, \mathbf{H_{cs}});\mathbf{F^i_{h'}}])$. Finally, each fact encoding is average pooled and concatenated with the predicted intent logits I_{tgt}, followed by a fully connected neural network f_{k_pred} to predict relevancy $F^i_{pred} = f_{k\_pred}([\text{avg}(\mathbf{F^i_h});I_{tgt}])$, which is trained by minimizing the loss between the prediction and the true label.

5.2 Decoder

Apart from the hidden encoder representation $\mathbf{C_{emb}}$, we also condition the response generation on the response personality and intent control codes, which enables the model to adapt to the required style. Similar to Rashkin et al. (2021), the control codes are prepended to the decoder input ids, and passed to the decoder, which generates the response by conditioning on the encoder context $\mathbf{C_{emb}}$, and the control codes. The entire system is trained end-to-end by minimizing the weighted sum of the language modelling cross entropy loss, the binary cross entropy fact selection loss, binary cross entropy intent prediction loss, and the cross entropy trait prediction loss.

6 Experiments and Results

6.1 Experiment Set-up

We used the pre-trained 139M parameters (base) version of BART (Lewis et al., 2020), and the 400M parameters distilled BlenderBot (Roller et al., 2020) from the Huggingface library (Wolf et al., 2020) as our base models, and added 24 new tokens comprising of speaker identifiers (agent_1, agent_2), traits and intent control codes to the embedding layer. Similar to Transfertransfo (Wolf et al., 2019), we introduce a token type embedding layer to demarcate turns. We utilized a learning rate of 2E-5, and batch size of 32 and 16 per GPU for BART and BlenderBot respectively, with gradient accumulation (Lin et al., 2018) for 2 steps, for BlenderBot. We clipped (Pascanu et al., 2013) the gradients to unit norm, and used AdamW (Loshchilov and Hutter, 2019) with default PyTorch parameters for optimization. Beam search was used during decoding with a beam length of 5, with penalty for trigram repetitions within the generated text, and between the context and generated text. The corpus based codes are only input to the encoder to aid in trait and intent predictions, and are not used as stylistic control codes.

Corpus	Model	Perplexity	BLEU 4	Rouge L	BLEURT
WOW	E2E (Dinan et al., 2019)	23.1/32.8	1.5 / 0.3		
	GPT2 (Rashkin et al., 2021)		8.9 / 8.4		
	T5 (Rashkin et al., 2021)		8.4 / 8.7		
	BART	9.74 / 10.53	8.44 / 8.24	0.341 / 0.342	0.491 / 0.488
	BART + All (P-Traits)	**9.37 / 10.13**	9.01 / 8.60	**0.349 / 0.349**	0.502 / **0.502**
	BART + All (E-Traits)	9.43 / 10.23	**9.20 / 8.79**	0.348 / 0.347	**0.506** / 0.501
	BlenderBot	7.48 / 8.54	**6.31** / 4.77	0.302 / 0.282	**0.462 / 0.444**
	BlenderBot + All (P-Traits)	7.38 / 8.39	6.22 / **4.90**	**0.305 / 0.294**	0.450 / 0.437
	BlenderBot + All (E-Traits)	**7.37 / 8.38**	6.22 / 4.77	0.304 / 0.294	0.451 / 0.441
TC	NRG (Gopalakrishnan et al., 2019)	26.30 / 36.30			
	PD-NRG (Hedayatnia et al., 2020)	12.25 / 12.62	1.9 / 2.0	0.113 / 0.108	
	Proto (Saha et al., 2021)	11.55 / 10.87			
	BART	13.81 / 14.71	3.62 / 4.10	0.235 / 0.250	0.365 / 0.388
	BART + All (P-Traits)	**13.21** / 14.10	3.72 / **4.37**	0.242 / **0.259**	0.370 / 0.400
	BART + All (E-Traits)	13.22 / **14.02**	**3.73** / 4.28	**0.246** / 0.258	**0.376 / 0.403**
	BlenderBot	11.09 / 10.75	3.13 / **3.75**	0.223 / 0.240	0.367 / 0.390
	BlenderBot + All (P-Traits)	10.75 / 10.39	**3.22** / 3.65	0.232 / 0.247	0.367 / 0.389
	BlenderBot + All (E-Traits)	**10.72 / 10.35**	3.20 / 3.62	**0.234 / 0.247**	**0.369 / 0.391**

Table 2: Language modelling results on the seen/unseen and frequent/rare topic portions of WOW and TC test sets.

6.2 Evaluating Language Modelling

For automatically evaluating the language modelling capabilities of our proposed model we compute and compare language modelling perplexity, BLEU-4 (Papineni et al., 2002) and ROUGE-L (Lin, 2004) scores. Since BLEU and ROUGE are known to be incomplete metrics, as they don't completely capture sentence semantics, we also compare the BLEURT (Sellam et al., 2020) scores. We report our results and compare with baselines in Table 2. For both WOW and TC, we consider the models void of any control codes, using only conversation context and facts as the internal baseline (underlined), and compare against variations containing both the Pandora and Essays based personality and intent based control codes, All (P-Traits), and All (E-Traits) respectively. As reference, we also include results of the end-to-end generative model (E2E) with gold knowledge that was introduced in the original WOW paper (Dinan et al., 2019), and the GPT-2 and T5 based knowledge grounded models proposed by (Rashkin et al., 2021) for WOW. For TC, we include results from the neural response generator (NRG) model introduced in the original paper (Gopalakrishnan et al., 2019), the follow up work using policy driven approach (PD-NRG) (Hedayatnia et al., 2020), and the recent work by Proto (Saha et al., 2021). For each dataset and model type in Table 2, we highlight in bold the best performing model by each metric, and underline the metric wise best performing models for a dataset. We observe: (i) In comparison to both the internal and external baselines, conditioning on intent and personality trait based control codes consistently yields better automatic evaluation scores. We reason that the introduction of control codes not only provides additional supervision signals, but also helps the language model to better factorize the probability distributions of the words. (ii) Using BlenderBot yields better perplexity scores, at the cost of precision/recall based metrics. We reason that although the extensive pre-training of BlenderBot on the BST dataset (Smith et al., 2020) helps in language modelling, its low vocabulary size of 8,008 tokens compared to 50,265 of BART, hinders adapting to the new datasets. (iii) Both the Essays and Pandora based codes work well; The Pandora based codes seem to work slightly better for WOW, while the Essays based codes perform better for TC. We reason that as depicted in Table 12, the Pandora based personality classifier identifies more instances of openness compared to the Essays based classifier. Since being objective is associated with the trait of openness, and the WOW dataset has 71% objective exchanges, which is more compared to 51% in the TC dataset (Table 7), it works better for WOW.

6.3 Evaluating Stylistic Control

We introduce two automatic metrics for comparing the intent and personality traits exhibited by the generated response and the golden response: (i) **Intent F1:** Re-using the intent classifier used for automatic annotation from section 4.2.2, we predict the intents exhibited by each of the the generated responses, and calculate the F1 score between the exhibited intents and the actual desired intent. We further derive a single metric by averaging the F1 score for all classes. (ii) **Trait Correlation:** Reusing the Big-5 personality trait intensity prediction models from section 4.2.2, we predict the intensity of each trait exhibited by the generated response, and compute the Pearson's correlation between the actual intensity from the golden response. We further average the correlation score across all the 5 traits to derive a single metric. Table 3 reports the results; For each dataset and model type we highlight the best performing model by each metric in bold, and underline the metric wise best performing models for each dataset. We observe: (i) Models that utilize the stylistic control codes during response generation yield better results, compared to the baseline versions which don't use any control codes. This indicates the effectiveness of our proposed method of controlling the response generation using stylistic control codes. (ii) Compared to using Pandora based personality codes, responses from models incorporating the Essays based control codes correlate more to the desired response trait. (iii) In majority cases, responses from models incorporating the Pandora and intent based control codes confirm more to the desired response intent, compared to models using the Essays based control codes along with intent for controlling personality. This hints towards possible synergic relationships between the personality and intent based codes.

Corpus	Model	Intent F1	Trait Correl.
WOW	BART	0.300 / 0.319	0.850 / 0.824
	BART + All (P-Traits)	**0.669 / 0.683**	0.858 / 0.836
	BART + All (E-Traits)	0.634 / 0.639	**0.870 / 0.848**
	BlenderBot	0.316 / 0.321	0.825 / 0.804
	BlenderBot + All (P-Traits)	0.466 / 0.469	0.828 / 0.810
	BlenderBot + All (E-Traits)	**0.480 / 0.491**	**0.835 / 0.818**
TC	BART	0.264 / 0.256	0.726 / 0.763
	BART + All (P-Traits)	**0.505 / 0.523**	0.731 / 0.765
	BART + All (E-Traits)	0.465 / 0.468	**0.748 / 0.782**
	BlenderBot	0.267 / 0.261	0.691 / 0.733
	BlenderBot + All (P-Traits)	**0.518 / 0.517**	0.720 / 0.749
	BlenderBot + All (E-Traits)	0.517 / 0.513	**0.737 / 0.768**

Table 3: Stylistic control results on the seen/unseen and frequent/rare topic portions of WOW and TC.

6.4 Ablation Study

We further perform the following ablation study with diverse combinations of the stylistic control codes for observing the effect of each type of code independently: (i) Intent: Using only intent based control codes in the decoder. (ii) C-Traits: Using only corpus based traits in the encoder, without any control codes in the decoder. (iii) P / E-Traits: Using only Pandora or Essays based personality control codes in the decoder. (iv) Intent + P / E-Traits: Using both intent and personality control codes in the decoder. (v) All: Using both intent and personality control codes in the decoder, along with corpus traits in the encoder. Table 9 reports the results of the ablation study. We observe: (i) Using intent as stylistic control code mostly yields better results for all metrics, compared to the baseline. (ii) Leveraging the corpus traits in the encoder alone, without incorporating any control codes in the decoder mostly yields poor results for all metrics. (iii) Incorporating both intent and personality codes in the decoder mostly yields best results across all metrics. The addition of corpus based traits in the encoder further improves the results in some cases.

6.5 Human Evaluation for Relevance

Since the ultimate desired impact of our model is to find use in human facing systems, we further perform human evaluation to qualitatively compare the responses generate by the trained models. Considering only the BART based models as they resulted in better BLEU and ROUGE scores in table 2, we sampled 100 examples from the test sets of each dataset, and utilized Amazon Mechanical Turk for performing human evaluation. An initial manual evaluation by us revealed that almost all the generated candidates were fluent, which can be attributed to the extensive pre-training of the state-of-the-art base language models. This is further corroborated by the human evaluations performed by Rashkin et al. (2021), where they observed fluency to be the least dimension of variation. Hence, employing 2 evlauators per sample, we evaluated the relevance of each candidate responses to the conversation context on a Likert scale from 1 to 5, where a rating of 1 signifies low relevance and 5 signifies high relevance. More details about the annotation framework is presented in the appendix A.4. Table 5 includes the averaged results from the human evaluation. We highlight the best scoring model per dataset in bold, and perform Welch's t-test to mark models which perform significantly lower than the best score with asterisks. Further, we calculate the inter annotator agreement using Krippendorff's alpha and find that the agreement to be > 0.90, indicating a high agreement. The results indicate that responses generated by controlling the personality traits were deemed more relevant to the conversation. Additionally, incorporating intent based stylistic codes along with personality achieves best results for the WOW dataset.

ID	Conversation Context	Generated Stylistic Response	Generated Variation Response	Style
1	**Agent 1**: Fred missed the penalty, and was dropped from the team. I wonder what he is going through.	i'm sorry to hear that. i'm sure he's going through some tough times. i hope he's ok.	i'm not sure what he's going through right now. i'm sure he is struggling.	Strong Agr
2	**Agent 1**: do you want to go on a hike this weekend? **Agent 2**: sure. where are we going? **Agent 1**: how about Yosemite?	yosemite national park. it's surrounded on the southeast by the Sierra national forest and on the northwest by the Stanislaus national forest. i've never been there.	yosemite is a beautiful place. i would love to hike there.	Objectivity
3	**Agent 1**: do you like to party?	i love to party! do you?	i do not. do you?	Strong Agr Strong Ext Weak Neu
4	**Agent 1**: how's it going, do you follow football? depending on the country it could refer to rugby or soccer.	i do follow football. it's one of the four major professional sports leagues in north america and the highest professional level of american football in the world. how about you?	i don't really follow football. i do know that the nfl is one of the four major professional sports leagues in north america and the highest professional level of american football in the world.	Strong Agr Strong Ext Weak Neu
5	**Agent 2**: i know! and brady won the super bowl in the first year he was a starter! **Agent 1**: wasn't the only other quarterback to do that kurt warner?	yes, he was! i think he was the first to win a super bowl. i wonder if he ever played in the nfl? did you know those huge american flags at football stadiums weigh 1,100 pounds and cost $50,000 to make?	no, i'm not sure, but i do know that those huge american flags at football stadiums weigh 1,100 pounds and cost $50,000 to make!	Strong Agr Strong Opn Strong Ext
6	**Agent 2**: congrats! that's awesome! when i fish in my local lake, i usually use lures instead of bait. it's easier and the fish seem to respond just as well. **Agent 1**: well we definitely got our money's worth, but it was so long just to get out there, and it was cold.	that's true. some hobbyists even make handmade tackle themselves, using plastic lures and artificial flies. that's pretty cool!	that's a shame! i'm glad you didn't have to use plastic lures or artificial flies to catch the red fish.	Strong Agr Strong Opn Strong Con Strong Ext Weak Neu Objectivity

Table 4: Generated samples with different combinations of stylistic control codes.

Model	TC	WOW
BART	3.54*	3.44**
BART + Intent	3.51*	3.61
BART + Big-5 Traits	**3.73**	3.58
BART + Intent + Big-5 Traits	3.47*	3.45**
BART + All	3.5**	**3.71**

Table 5: Human evaluation results: *, ** indicates that this result is significantly different from the best result in that column with p-value < 0.05 and < 0.02 respectively.

6.6 Discussion

Table 4 showcases a few style controlled responses generated by our proposed models. For each conversation context, we leverage the control code in the style column and generate the stylistic response. We further contrast the stylistic response against a variation response generated either using randomly selected control codes or the baseline model without any stylistic codes. Example 1 demonstrates how incorporating strong agreeableness as stylistic code results in the response exuding empathy, in comparison to the variation response. Example 2 demonstrates the model's capability of generating objective response, by leveraging external facts. Through examples 3-6 we demonstrate the model's capability of simultaneously incorporating multiple stylistic codes during generation. Examples 3-6 demonstrate how increasing agreeableness results in a positive stance in the response. We also notice in examples 4 and 5 how increasing extraversion results in the model asking open-ended questions, thus portraying an extroverted and outgoing personality. Further, in examples 4 and 6 we notice the effect of controlling neuroticism, where the variant response is not consistent compared to the stylistic controlled response. Overall, we observe that utilizing our proposed method, it is possible to control the style of the response using stylistic control codes, and further combine different codes to generate compounded stylistic responses.

7 Conclusion

Here we experiment with training end-to-end methods for controlling the response style in generative conversational models. We believe incorporating such methods in human facing dialogue systems should benefit the system by providing it with more control. Using combinations of Big-5 personality traits and dialogue intent based stylistic control codes during language modelling, we are able to successfully control the style of a response as desired, the efficacy of which is further established by the achieved results. Additionally, we engender two annotated dialogue corpora with intents and personality traits for use by the community.

References

Johnstone Barbara. 2017. *Discourse Analysis.*, volume Third edition of *Introducing Linguistics*. Wiley-Blackwell.

Ethan A. Chi, Chetanya Rastogi, Alexander Iyabor, Hari Sowrirajan, Avanika Narayan, and Ashwin Paranjape. 2021. Neural, neural everywhere: Controlled generation meets scaffolded, structured dialogue.

Paul T Costa Jr. 1992. Revised neo personality inventory and neo five-factor inventory. *Professional manual*.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *Proceedings of the 2019 Conference of the North*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Alan Dix, Janet Finlay, Gregory D Abowd, and Russell Beale. 2003. *Human-computer interaction*. Pearson Education.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Matej Gjurković, Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Snajder. 2021. PANDORA talks: Personality and demographics on Reddit. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 138–152, Online. Association for Computational Linguistics.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Vrindavan Harrison, Lena Reed, Shereen Oraby, and Marilyn Walker. 2019. Maximizing stylistic control and semantic accuracy in NLG: Personality variation and discourse contrast. In *Proceedings of the 1st Workshop on Discourse Structure in Neural NLG*, pages 1–12, Tokyo, Japan. Association for Computational Linguistics.

Behnam Hedayatnia, Karthik Gopalakrishnan, Seokhwan Kim, Yang Liu, Mihail Eric, and Dilek Hakkani-Tur. 2020. Policy-driven neural response generation for knowledge-grounded dialogue systems.

Alexei V. Ivanov, Giuseppe Riccardi, Adam J. Sporka, and Jakub Franc. 2011. Recognition of personality traits from human spoken conversations. In *INTERSPEECH*.

Alexandra Jaffe et al. 2009. *Stance: sociolinguistic perspectives*. Oup Usa.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*.

Jakub Konrád, Jan Pichl, Petr Marek, Petr Lorenc, Van Duy Ta, Ondřej Kobza, Lenka Hýlová, and Jan Šedivý. 2021. Alquist 4.0: Towards social intelligence using generative models and dialogue personalization.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yujun Lin, Song Han, Huizi Mao, Yu Wang, and Bill Dally. 2018. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Andrea Madotto, Etsuko Ishii, Zhaojiang Lin, Sumanth Dathathri, and Pascale Fung. 2020. Plug-and-play conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2422–2433.

François Mairesse and Marilyn Walker. 2007. Personage: Personality generation for dialogue. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 496–503.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.

François Mairesse and Marilyn A. Walker. 2008a. A personality-based framework for utterance generation in dialogue applications. In *AAAI Spring Symposium: Emotion, Personality, and Social Behavior*.

François Mairesse and Marilyn A. Walker. 2008b. Trainable generation of big-five personality styles through data-driven parameter estimation. In *ACL*.

Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2017. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2):74–79.

Isabel Briggs Myers. 1962. The myers-briggs type indicator: Manual (1962).

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1699–1713, Online. Association for Computational Linguistics.

Shereen Oraby, Lena Reed, Shubhangi Tandon, Sharath T.S., Stephanie Lukin, and Marilyn Walker. 2018. Controlling personality-based stylistic variation with neural natural language generators. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 180–190, Melbourne, Australia. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *ICML*.

James Pennebaker and Laura King. 2000. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77:1296–312.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 704–718, Online. Association for Computational Linguistics.

Zhancheng Ren, Qiang Shen, Xiaolei Diao, and Hao Xu. 2021. A sentiment-aware deep learning approach for personality detection from text. *Information Processing Management*, 58(3):102532.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot.

Sougata Saha, Souvik Das, Elizabeth Soper, Erin Pacquetet, and Rohini K. Srihari. 2021. Proto: A neural cocktail for generating appealing conversations.

Fillmore H Sanford. 1942. Speech and personality. *Psychological Bulletin*, 39(10):811.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

Christopher Soto. 2018. *Big Five personality traits*, pages 240–241.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

WU Wanqi and Tetsuya Sakai. 2020. Response generation based on the big five personality traits.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents.

Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation.

A Appendix

A.1 Big-5 Personality Trait Annotation

We utilized the Pandora and Essays datasets to train automatic personality predictors. The Pandora dataset consists of multiple Reddit posts for a user, along with the actual Big-5 trait intensities for the user, whereas the Essays dataset consist of essays written by psychology students, with actual Big-5 trait labels, which we converted to intensities, for maintaining parity between the datasets. For both the datasets, we tokenized the text into sentences, and maintained a list of sentences for each user. We further cleansed and normalised the sentence lists, and preserved sentences containing ASCII characters with 3 to 50 tokens. In order to make the length distribution of the training examples similar to conversation datasets, for each user we derived m non-overlapping samples by randomly selecting and concatenating k sentences, where k was randomly selected to vary between 2 and 30. The target intensities for each of the Big-5 traits were kept same for the m samples, and were scaled to vary between -1 and 1. Overall, we derived 7,230 train and 804 validation examples from Essays, and 75,172 training, and 39,447 validation examples from the Pandora dataset. We incorporated fully connected layers followed by Tanh activation on top of RoBERTa base, to predict all the 5 trait intensities simultaneously, and trained the models to minimize mean squared error loss. With the intention of comparing the quality and usefulness of the automatic personality annotations, we trained 2 versions of the models, one for each personality dataset. In order to leverage pre-training, the model trained on Essays dataset was initialized from a checkpoint of the Pandora model. Both the models were trained with a batch size of 32, and learning rate of 2E-5, till validation loss ceased improving. We leveraged AdamW optimizer for optimizing the model parameters, and resorted to mixed precision training to reduce the training time. In Table 6, for each trait we report Pearson correlation between the predicted intensity and the actual values for both the datasets. Using 0 as a threshold, we further binarize the predicted intensities and actual labels, and report classification F1.

Trait	Essays Pearson Correl.	Essays F1	Pandora Pearson Correl.	Pandora F1
Agr	0.228	0.640	0.813	0.832
Opn	0.321	0.620	0.813	0.902
Con	0.276	0.578	0.797	0.776
Ext	0.255	0.568	0.808	0.799
Neu	0.249	0.658	0.799	0.848

Table 6: Correlation and F1 metrics on the respective validation dataset for the Pandora based and Essays based model.

A.2 Fact Selection Example Creation

During fact selection, for both the Topical Chat and Wizard of Wikipedia we presented 5 external facts per example to choose from, for each interlocutor turn. The 5 facts comprised of the golden fact(s) required for generating the current response, and the remaining were randomly sampled from the facts which are available to the interlocutor. Table 7 contains the percentage distribution of the positive class for fact selection, and for each dialogue intent.

Corpus	Split	Subj	Obj	Subj Q	Obj Q	Fact
WOW	Seen	46%	71%	6%	2%	18%
WOW	Unseen	43%	71%	6%	2%	18%
TC	Frequent	68%	51%	12%	6%	5%
TC	Rare	70%	52%	13%	4%	7%

Table 7: Percentage distribution of positive class for each intent type, and fact selection in Wizard of Wikpedia and Topical Chat.

A.3 Additional Results

Table 8 reports the F1 scores of the best performing models for predicting each of the additional tasks in the multi-task learning framework. Table 9

Type	Model (WOW)	F1 (WOW)	Model (TC)	F1 (TC)
Fact	BART + Intent / BART + All (P-Traits)	0.50 / 0.44	BlenderBot / BlenderBot	0.13 / 0.12
Subj	BART + All (E-Traits) / BART + All (E-Traits)	0.75 / 0.73	BART + All (E-Traits) / BART + Intent + E-Traits	0.83 / 0.84
Obj	BART + All (P-Traits) / BART + All (P-Traits)	0.86 / 0.86	BART + Intent / BlenderBot + All (E-Traits)	0.69 / 0.70
Subj Q	BlenderBot + E-Traits / BlenderBot + E-Traits	0.58 / 0.59	BART + E-Traits / BART + E-Traits	0.63 / 0.63
Obj Q	BlenderBot + E-Traits / BlenderBot + E-Traits	0.58 / 0.60	BART + E-Traits / BART + E-Traits	0.61 / 0.64
Agr	BART + All (E-Traits) / BART + All (E-Traits)	0.61 / 0.58	BART + Intent + E-Traits / BART + E-Traits	0.64 / 0.66
Opn	BART + All (E-Traits) / BART + All (E-Traits)	0.46 / 0.44	BlenderBot + All (E-Traits) / BlenderBot + All (E-Traits)	0.47 / 0.46
Con	BART + All (E-Traits) / BART + All (E-Traits)	0.61 / 0.62	BART + Intent + E-Traits / BART + Intent + E-Traits	0.63 / 0.63
Ext	BART + All (E-Traits) / BART + All (E-Traits)	0.61 / 0.62	BART + All (E-Traits) / BART + Intent + E-Traits	0.62 / 0.65
Neu	BART + All (E-Traits) / BART + All (E-Traits)	0.62 / 0.61	BART + Intent + E-Traits / BART + All (E-Traits)	0.61 / 0.66

Table 8: F1 scores of the best performing planning models for each policy component, in both the seen/unseen splits of Wizard of Wikipedia (WOW), and frequent/rare splits of Topical Chat (TC) test sets.

Corpus	**Model**	**Perplexity**	**BLEU 4**	**RougeL**	**BLEURT**	**Intent F1**	**Trait Correl.**
WOW	BART	9.74 / 10.53	8.44 / 8.24	0.341 / 0.342	0.491 / 0.488	0.300 / 0.319	0.85 / 0.824
	BART + Intent	9.43 / 10.23	8.69 / 7.96	0.338 / 0.335	0.495 / 0.492	0.469 / 0.486	0.848 / 0.824
	BART + C-Traits	9.76 / 10.52	8.32 / 8.11	0.338 / 0.338	0.487 / 0.486	0.297 / 0.300	0.849 / 0.826
	BART + P-Traits	9.53 / 10.27	8.72 / 8.45	0.344 / 0.347	0.496 / 0.492	0.402 / 0.406	0.855 / 0.827
	BART + E-Traits	9.52 / 10.27	8.99 / 8.58	0.345 / **0.349**	0.496 / 0.494	0.395 / 0.397	0.866 / 0.844
	BART + Intent + P-Traits	9.41 / 10.21	9.22 / 8.44	0.345 / 0.342	0.502 / 0.496	0.618 / 0.636	0.856 / 0.833
	BART + Intent + E-Traits	**9.37** / 10.14	**9.25** / 8.51	0.346 / 0.345	0.502 / 0.500	0.654 / 0.656	0.866 / **0.849**
	BART + All (P-Traits)	**9.37 / 10.13**	9.01 / 8.60	**0.349** / **0.349**	0.502 / **0.502**	**0.669** / **0.683**	0.858 / 0.836
	BART + All (E-Traits)	9.43 / 10.23	9.20 / **8.79**	0.348 / 0.347	**0.506** / 0.501	0.634 / 0.639	**0.870** / 0.848
	BlenderBot	7.48 / 8.54	6.31 / 4.77	0.302 / 0.282	**0.462** / 0.444	0.316 / 0.321	0.825 / 0.804
	BlenderBot + Intent	**7.35** / 8.38	**6.52 / 5.29**	**0.311 / 0.297**	**0.462** / **0.449**	**0.570** / **0.564**	0.834 / 0.809
	BlenderBot + C-Traits	7.49 / 8.54	6.33 / 5.00	0.301 / 0.286	0.460 / 0.447	0.320 / 0.329	0.825 / 0.801
	BlenderBot + P-Traits	7.42 / 8.44	6.24 / 4.90	0.306 / 0.293	0.456 / 0.445	0.369 / 0.370	0.831 / 0.809
	BlenderBot + E-Traits	7.41 / 8.42	6.37 / 4.89	0.309 / 0.293	0.459 / 0.445	0.359 / 0.369	0.840 / **0.818**
	BlenderBot + Intent + P-Traits	7.37 / 8.38	6.26 / 5.01	0.307 / 0.295	0.455 / 0.442	0.472 / 0.485	0.833 / 0.811
	BlenderBot + Intent + E-Traits	7.36 / **8.37**	6.29 / 5.04	0.308 / 0.295	0.457 / 0.444	0.508 / 0.500	**0.841** / 0.817
	BlenderBot + All (P-Traits)	7.38 / 8.39	6.22 / 4.90	0.305 / 0.294	0.450 / 0.437	0.466 / 0.469	0.828 / 0.810
	BlenderBot + All (E-Traits)	7.37 / 8.38	6.22 / 4.77	0.304 / 0.294	0.451 / 0.441	0.480 / 0.491	0.835 / **0.818**
TC	BART	13.81 / 14.71	3.62 / 4.10	0.235 / 0.250	0.365 / 0.388	0.264 / 0.256	0.726 / 0.763
	BART + Intent	13.25 / 14.12	3.62 / 4.30	0.234 / 0.251	0.373 / 0.399	0.359 / 0.377	0.723 / 0.767
	BART + C-Traits	13.73 / 14.68	3.49 / 4.13	0.233 / 0.251	0.361 / 0.390	0.263 / 0.267	0.725 / 0.759
	BART + P-Traits	13.59 / 14.57	3.60 / 4.12	0.236 / 0.253	0.363 / 0.390	0.286 / 0.317	0.731 / 0.766
	BART + E-Traits	13.57 / 14.53	3.52 / 4.08	0.237 / 0.252	0.364 / 0.390	0.290 / 0.299	0.733 / 0.771
	BART + Intent + P-Traits	13.25 / 14.14	3.69 / 4.20	0.239 / 0.252	0.364 / 0.392	0.461 / 0.471	0.729 / 0.773
	BART + Intent + E-Traits	**13.21** / 14.10	**3.75** / **4.38**	**0.246** / **0.259**	**0.377** / **0.403**	0.459 / 0.470	0.747 / **0.783**
	BART + All (P-Traits)	**13.21** / 14.10	3.72 / 4.37	0.242 / **0.259**	0.370 / 0.400	**0.505 / 0.523**	0.731 / 0.765
	BART + All (E-Traits)	13.22 / **14.02**	3.73 / 4.28	**0.246** / 0.258	0.376 / **0.403**	0.465 / 0.468	**0.748** / 0.782
	BlenderBot	11.09 / 10.75	3.13 / 3.75	0.223 / 0.240	0.367 / 0.390	0.267 / 0.261	0.691 / 0.733
	BlenderBot + Intent	10.79 / 10.45	**3.41 / 3.85**	0.230 / **0.247**	**0.373 / 0.396**	0.472 / 0.480	0.713 / 0.747
	BlenderBot + C-Traits	11.09 / 10.75	3.22 / 3.75	0.222 / 0.240	0.365 / 0.390	0.273 / 0.268	0.695 / 0.737
	BlenderBot + P-Traits	11.01 / 10.65	3.16 / 3.66	0.227 / 0.243	0.366 / 0.390	0.326 / 0.336	0.710 / 0.745
	BlenderBot + E-Traits	10.98 / 10.61	3.18 / 3.66	0.229 / 0.246	0.369 / 0.391	0.329 / 0.334	0.732 / 0.766
	BlenderBot + Intent + P-Traits	10.76 / 10.41	3.19 / 3.64	0.232 / **0.247**	0.368 / 0.390	**0.524** / **0.531**	0.715 / 0.753
	BlenderBot + Intent + E-Traits	10.73 / 10.37	3.13 / 3.66	**0.234 / 0.247**	0.370 / 0.392	0.513 / 0.525	0.733 / **0.770**
	BlenderBot + All (P-Traits)	10.75 / 10.39	3.22 / 3.65	0.232 / **0.247**	0.367 / 0.389	0.518 / 0.517	0.720 / 0.749
	BlenderBot + All (E-Traits)	**10.72** / **10.35**	3.20 / 3.62	**0.234 / 0.247**	0.369 / 0.391	0.517 / 0.513	**0.737** / 0.768

Table 9: Ablation study on the seen/unseen and frequent/rare topic portions of the Wizard of Wikipedia (WOW), and Topical Chat (TC) test sets. Best performing models are highlighted in bold.

contains the ablation study results. For each conversation corpus, and personality dataset combination, Table 12 lists the percentage distribution of strong and weak categories (seperated by '/') for each Big-5 trait, by each split of the dataset. Table 13 contains results without access to the golden policy consisting of control codes during inference. The model leverages the predicted control codes as policy for response generation. Figure 3 plots the context length wise style adaptation of the generated response, which hints lengthier context facilitates better adaptation to the desired response style.

A.4 Amazon Mechanical Turk for Evaluation

We leveraged Amazon Mechanical Turk (AMT) in order to perform human evaluations on our model

Intent	Percentage Occurance	Annotator Agreement
State Knowledge Fact	0.33	0.95
State Opinion	0.30	0.81
State Personal Fact	0.13	1.00
Request (Opinion/Knowledge Fact/Personal Fact)	0.13	1.00
Others	0.11	1.00

Table 10: Intent Automatic Annotation Evaluation.

Personality Trait	Pandora Occurance	Pandora Agreement	Essays Occurance	Essays Agreement
agreeableness	0.195	0.696	0.197	0.760
openness	0.229	0.630	0.252	0.563
conscientiousness	0.169	0.550	0.181	0.391
extraversion	0.186	0.364	0.189	0.750
neuroticism	0.220	0.308	0.181	0.435

Table 11: Personality Trait Automatic Annotation Evaluation.

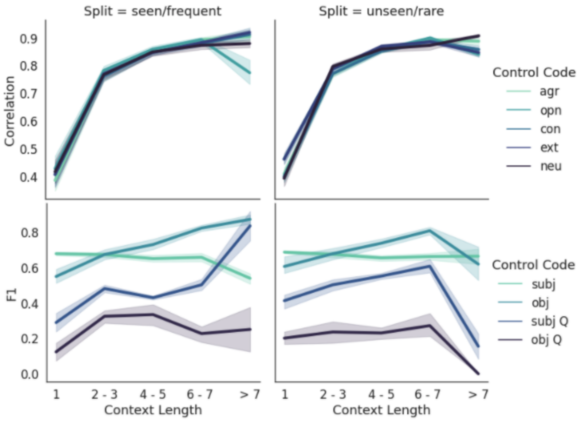


Figure 3: Turn length wise adaptation to the desired response style, collated from all the full version models.

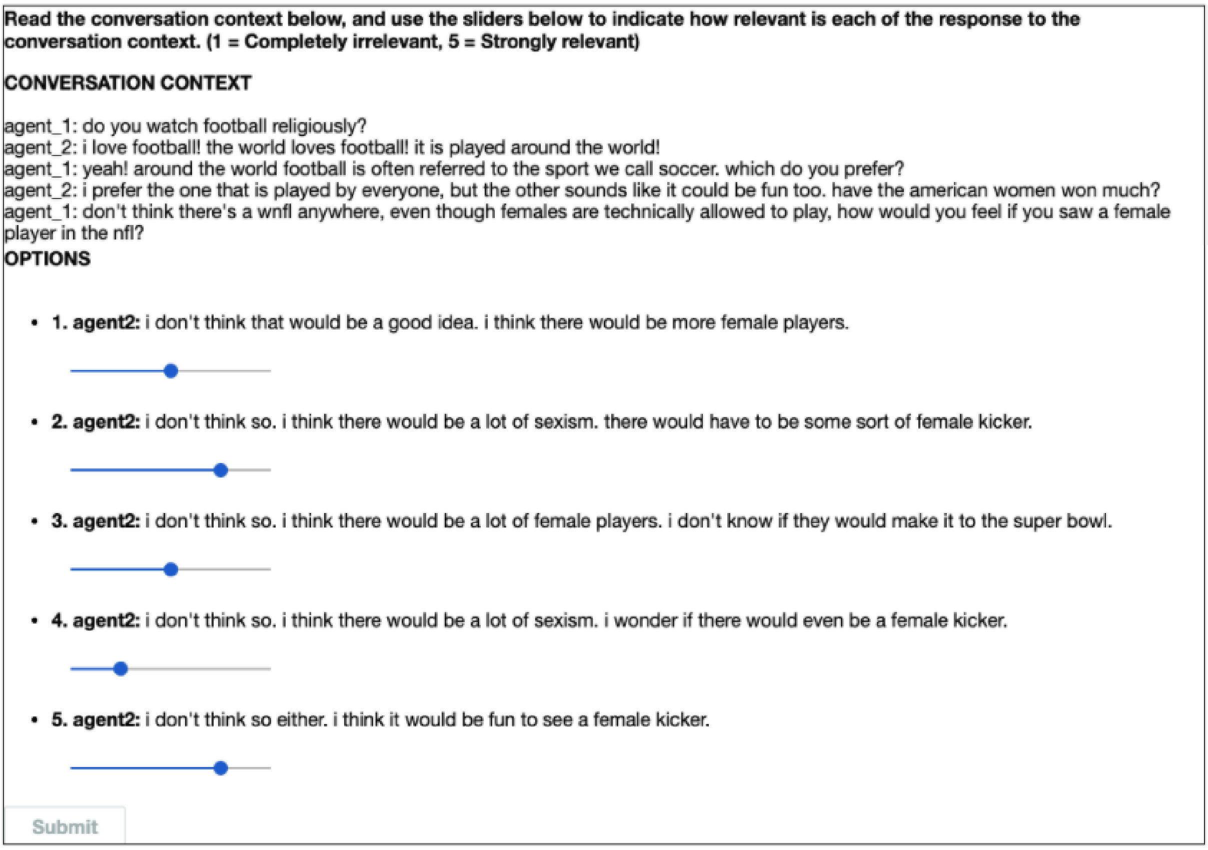
response. We set up human intelligence task (HIT) in the AMT platform, with two evaluators per example and each task worth $0.01. The evaluators were provided with clear instructions on what to annotate and how to annotate the examples. The task comprised of reading a conversation context, and rating 5 different responses on a Likert scale of 1 to 5, where the responses were generated by different models, unknown to the annotator. Figure 4 illustrates a sample screenshot of the HIT interface along with the instructions used for collecting the evaluations.

		Seen/ Frequent Topic					Unseen/ Rare Topic				
Corpus	Personality Corpus	Agr	Opn	Con	Ext	Neu	Agr	Opn	Con	Ext	Neu
WOW	Pandora	19/20	80/8	19/19	17/20	19/20	20/18	81/8	17/20	12/24	22/15
	Essays	22/15	78/10	20/17	21/15	16/20	21/12	79/10	15/18	20/16	14/20
TC	Pandora	47/18	72/10	29/25	39/19	20/33	20/38	67/16	22/37	12/46	37/18
	Essays	40/12	61/23	38/14	49/8	7/49	22/29	65/17	14/41	11/45	40/17

Table 12: Percentage of Strong/Weak categories for all traits in each chat corpus, split by each personality corpus.

Corpus	Model	BLEU 4	RougeL	BLEURT
WOW	BART	8.44 / 8.24	0.341 / 0.342	0.491 / 0.488
	BART + Intent	8.63 / 7.87	0.334 / 0.332	0.495 / 0.491
	BART + C-Traits	8.32 / 8.11	0.338 / 0.338	0.487 / 0.486
	BART + P-Traits	8.69 / 8.42	0.343 / 0.342	0.494 / 0.489
	BART + E-Traits	8.94 / 8.60	0.342 / 0.344	0.495 / 0.490
	BART + Intent + P-Traits	9.41 / 8.47	0.342 / 0.336	0.499 / 0.490
	BART + Intent + E-Traits	8.86 / 8.12	0.337 / 0.332	0.497 / 0.491
	BART + All (P-Traits)	9.09 / 8.60	0.343 / 0.343	0.496 / 0.498
	BART + All (E-Traits)	9.26 / 8.82	0.340 / 0.343	0.499 / 0.495
	BlenderBot	6.31 / 4.77	0.302 / 0.282	0.462 / 0.444
	BlenderBot + Intent	6.36 / 5.20	0.301 / 0.287	0.457 / 0.446
	BlenderBot + C-Traits	6.33 / 5.00	0.301 / 0.286	0.460 / 0.447
	BlenderBot + P-Traits	6.28 / 4.98	0.306 / 0.289	0.453 / 0.441
	BlenderBot + E-Traits	6.34 / 4.90	0.305 / 0.288	0.457 / 0.441
	BlenderBot + Intent + P-Traits	6.32 / 4.99	0.301 / 0.289	0.450 / 0.440
	BlenderBot + Intent + E-Traits	6.21 / 4.99	0.300 / 0.288	0.452 / 0.441
	BlenderBot + All (P-Traits)	6.29 / 4.75	0.301 / 0.287	0.443 / 0.430
	BlenderBot + All (E-Traits)	6.18 / 4.77	0.299 / 0.286	0.448 / 0.433
TC	BART	3.62 / 4.10	0.235 / 0.250	0.365 / 0.388
	BART + Intent	3.40 / 4.00	0.228 / 0.243	0.369 / 0.397
	BART + C-Traits	3.49 / 4.13	0.233 / 0.251	0.361 / 0.390
	BART + P-Traits	3.54 / 4.10	0.233 / 0.250	0.362 / 0.389
	BART + E-Traits	3.40 / 4.01	0.233 / 0.248	0.363 / 0.388
	BART + Intent + P-Traits	3.32 / 3.92	0.227 / 0.240	0.361 / 0.389
	BART + Intent + E-Traits	3.29 / 4.00	0.229 / 0.243	0.371 / 0.397
	BART + All (P-Traits)	3.36 / 3.96	0.227 / 0.242	0.366 / 0.396
	BART + All (E-Traits)	3.54 / 4.14	0.231 / 0.245	0.372 / 0.397
	BlenderBot	3.13 / 3.75	0.223 / 0.240	0.367 / 0.390
	BlenderBot + Intent	3.12 / 3.73	0.215 / 0.233	0.363 / 0.387
	BlenderBot + C-Traits	3.22 / 3.75	0.222 / 0.240	0.365 / 0.390
	BlenderBot + P-Traits	3.18 / 3.71	0.222 / 0.240	0.363 / 0.387
	BlenderBot + E-Traits	3.11 / 3.52	0.221 / 0.239	0.364 / 0.385
	BlenderBot + Intent + P-Traits	3.03 / 3.59	0.214 / 0.228	0.361 / 0.382
	BlenderBot + Intent + E-Traits	3.04 / 3.69	0.213 / 0.230	0.362 / 0.384
	BlenderBot + All (P-Traits)	3.06 / 3.50	0.214 / 0.229	0.359 / 0.382
	BlenderBot + All (E-Traits)	3.03 / 3.52	0.213 / 0.229	0.359 / 0.385

Table 13: Experimental results and ablation study on the seen/unseen and frequent/rare topic portions of the Wizard of Wikipedia (WOW), and Topical Chat (TC) test sets, using golden facts and model predicted stylistic control codes.



Read the conversation context below, and use the sliders below to indicate how relevant is each of the response to the conversation context. (1 = Completely irrelevant, 5 = Strongly relevant)

CONVERSATION CONTEXT

agent_1: do you watch football religiously?
agent_2: i love football! the world loves football! it is played around the world!
agent_1: yeah! around the world football is often referred to the sport we call soccer. which do you prefer?
agent_2: i prefer the one that is played by everyone, but the other sounds like it could be fun too. have the american women won much?
agent_1: don't think there's a wnfl anywhere, even though females are technically allowed to play, how would you feel if you saw a female player in the nfl?

OPTIONS

- **1. agent2:** i don't think that would be a good idea. i think there would be more female players.
- **2. agent2:** i don't think so. i think there would be a lot of sexism. there would have to be some sort of female kicker.
- **3. agent2:** i don't think so. i think there would be a lot of female players. i don't know if they would make it to the super bowl.
- **4. agent2:** i don't think so. i think there would be a lot of sexism. i wonder if there would even be a female kicker.
- **5. agent2:** i don't think so either. i think it would be fun to see a female kicker.

Figure 4: Sample screenshot from AMT HIT task.