Psych-E: Configurable Response Generation using Personality Traits and Pragmatics

Anonymous ACL submission

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 configurable response generators, in this paper we experiment with ways to ground neural response generators based on both (i) interlocutor Big-5 personality traits, and (ii) discourse intent as control codes, training an end-to-end dialogue agent that can not only leverage the control codes as policy for nuanced response generation, but also predict and decide the generation policy to be utilized by the generator. 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 policy consisting of noisy estimates of these features as control codes, and leverage automatic evaluation metrics along with ablation studies, to assess the impact of using control codes for response generation. Additionally, we leverage human judgement to demonstrate the effectiveness of using such personality and pragmatics based policy for response generation. Our experiments illustrate the effectiveness of this strategy resulting in improvements to existing benchmarks.

1 Introduction

Recent years have witnessed a growth in neural methods for language modelling (LM), 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). Although such models are capable of generating human-like responses, they come with their own set of predicaments. Leveraging only textual data, sans any other explicit control mechanism for training, such models often generate undesirable responses for a situation. (Rashkin et al., 2021) discuses the problem of knowledge hallucinations, (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 systems. In this paper, we experiment with ways to enhance the faithfulness of generated responses to the interlocutor personality traits, by leveraging personality and intent based control codes as response generation policy during training.

Figure 1: Sample dialogue between two users, depicting the influence of personality trait in speech.

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). Relying on language modelling (LM) for modelling dialogue without interlocutor specific supervision, might result in fluent, yet anomalous response. For example, the response to the query in Figure 1 is subjective, and dependent on the nature of the interlocutor. Had the interlocutor been introverted and exhibited weakly agreeable personality, the response could have been different. In order to factor in this phenomenon in LM, we incorporate Big 5 personality traits (Soto, 2018) as control codes, which is a well established personality taxonomy in psychological trait theory, and also one of the most recognized approaches to describe and measure individual differences in personality (Costa Jr, 1992). Employing automatic annotation schemes, we annotate 2 large scale open domain knowledge grounded chat corpora, and train end-to-end response generators which exhibit faithfulness to the interlocutor personality traits.

In a conversation, the speakers intentions shape the discourse. According to (Barbara, 2017), locutionary acts are equivalent to taking actions, and intentions are correlated with individual personality. Hence, we also experiment with leveraging pragmatics like dialogue intent as control codes for response generation. Further, treating intent and personality traits as generation policy, we experiment with leveraging the contextual policy, and the conversation history, in order to predict the target policy that should be followed by the current response. Thus, enabling a self sufficient system, that can predict the required response control codes, and incorporate the control cues while generating response.

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 and reliable large scale personality annotated text corpora. (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 three 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. 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. (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

Response Generation Our primary goal is to experiment with configurable response generation, using personality traits and dialogue intent as control codes for the decoder. We reason that since personality is the combination of behavior, emotion, motivation, and thought patterns that define an individual, conditioning response generation on such a feature can not only enable the model to factor in interlocutor idiosyncrasy during decoding, but also provide configurable knobs that can be used to vary the flavour of the response as needed. For our purpose, we utilize the Big-5 personality traits, along with corpus specific custom traits listed in

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.

Table 1 as control codes.

Dialogue intent, analogous to speech acts (Stolcke et al., 2000), elucidates the abstract or high level motives, and summarises the intention of a response. Hence, we reason that incorporating intent based control codes should not only enable response generation by apprising the high level meaning that the generated response should exude, but also provide us with additional configurations to regulate the response. For our use case, we re-purposed the intent taxonomy defined by (Saha et al., 2021), and derive four broad intent categories, described in Table 1. Overall, leveraging personality traits and intent as control codes, we not only provide the response generator with a policy for better modelling, but also provides us with a set of configurable parameters, that can be varied to generate diverse flavors of response.

Planning Conversation is considered as an interplay of conscious or subconscious interlocutor actions (Barbara, 2017), which arises from intent and personality. As a secondary goal, we experiment with leveraging the conversation context, and the historical interplay of personality traits and dialogue intent between interlocutors for predicting the target intent and trait prediction, which can be used as a generation policy consisting of control codes. We also experiment with empowering the model to select the most relevant fact from a set of input facts, which can be used for response generation, thus providing the model with control over the content that can be leveraged for generation. Overall, we experiment with training an end-to-end system that can not only plan the intent and traits (policy) to be exhibited by the response, but also decide the most relevant factual knowledge excerpts that can be leveraged by the generator, and generate a response that is faithful to the policy.

4 Data

We leverage publicly available, large scale Wizard of Wikipedia (Dinan et al., 2019), and Topical chat (Gopalakrishnan et al., 2019; Hedayatnia et al., 2020) corpora for our experiments. Both the datasets are multi-turn, knowledge grounded chat corpora. We further enrich the corpora with turn wise intent and personality trait annotations. Below we explain each of the datasets and the annotation scheme in detail.

4.1 Conversation Corpus

Wizard of Wikipedia (WOW): It is 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.

Topical Chat (TC): It is a more symmetric chat corpus consisting of conversations between two agents, where both the agents have access to diverse external knowledge sources. Compared to WOW, TC reflects real world conversations better, with lengthier conversations, and more subjectivity.

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, and further combine State Personal Fact and State

Opinion as Subj, Request Personal Fact and Request Opinion as Subj Q, while renaming State Knowledge Fact and Request Knowledge Fact to Obj and Obj Q respectively.

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 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. 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.

Corpus Based Traits We also define 3 interlocutor specific universal traits (Table 1), which are derived using corpus statistics. (i) **Attitude**: Captures the predominant stance (Jaffe et al., 2009) of an interlocutor 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 annotate 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 lead of 10% as the preferred tone, else both.(iii) **Length**: Captures whether an interlocutor prefers lengthy 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 and intent control codes P and I, along with relevant facts F and historical policy of the previous turns S, we model the posterior probability distribution $p(Y|C, P, I, F, S)$. Further, enabling response policy planning, we estimate $p(I|C, S)$, and $p(P|C, S)$, and for relevant fact selection F we estimate $p(F|C, S, I)$. The overall joint probability can be factorized as,

$$p(Y, C, P, I, F, S) = \prod_{i=1}^{n} p(y_i|y_{<i}, C, P, I, F, S) \\ p(I|C, S)p(P|C, S)p(F|C, S, I)p(C, S) \quad (313)$$

We employ parameterized neural networks to estimate each probability, 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 backbone architectures of our model. Figure 2 illustrates the end-to-end system, and below we detail each component [1].

5.1 Encoder

The encoder comprises of the encoding and the planning steps. It inputs the conversation context C, contextual policy S, and set of input facts F, and leverages 3 independent encoders and classifiers to output the encoded context representation $\mathbf{C_{emb}}$ for the decoder to use, along with the response policy comprising of the predicted personality traits P and intent I.

5.1.1 Encoding

The context encoder f_c encodes the context tokens C, and generates the representation $\mathbf{C_h}$. The policy representation $\mathbf{S_h}$ is obtained by aligning the

[1] The annotated dataset, models, and code to be made public on acceptance.

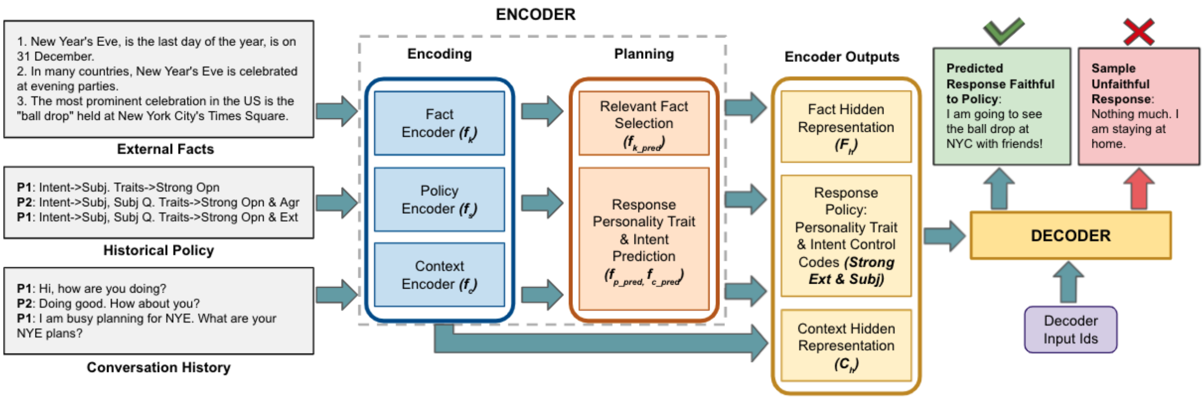

Figure 2: Proposed end-to-end system architecture for configurable, policy faithful response generation.

representation by the policy encoder f_s, with the context representation using multi-headed attention and feed forward layers $f_{s'}$. The steps can be summarized as:

$$\mathbf{C_h} = f_c(C),\ \mathbf{S_{h'}} = f_s(S)$$
$$\mathbf{S_h} = f_{s'}([\text{MultiHead}(\mathbf{S_{h'}}, \mathbf{C_h});\mathbf{S_{h'}}])$$
$$\mathbf{H_{cs}} = [\mathbf{C_h};\mathbf{S_h}],\ \vec{H_{cs}} = \text{avg}(\mathbf{H_{cs}})$$

For encoding the facts, we implement a fact encoder f_k for independently encoding each input fact F^i to the initial encoding $\mathbf{F^i_{h'}} = f_k(F^i)$. The final fact representation $\mathbf{F_h}$ is obtained by aligning each fact representation with the context and policy representation using multi-head attention and fully connected layers, followed by sum pooling the masked initial encoding $\mathbf{F^i_{h'}}$, where the mask is determines by the fact selector discussed below. This mechanism provides control to persist only the relevant fact representations for the decoder.

5.1.2 Planning

The planning module employs classifiers to predict the response policy, and also performs fact selection. We employ 2 fully connected neural networks: f_{i_pred} and f_{p_pred} to predict the response intent I and personality control codes P, which serve as the response policy. $I = f_{i\_pred}(\vec{H_{cs}})$, and $P = f_{p\_pred}(\vec{H_{cs}})$.

Deciding the most relevant fact from a set of external facts depends not only 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. Hence, for fact selection we align the fact encodings $\mathbf{F^i_{h'}}$ with the context and policy representations as $\mathbf{F^i_h} = f_{k'}([\text{MultiHead}(\mathbf{F^i_{h'}}, \mathbf{H_{cs}});\mathbf{F^i_{h'}}])$, and concatenate the predicted intent logits I with the average pooled fact encoding $\mathbf{F^i_h}$ for each fact, followed by a fully connected neural network f_{k_pred} to predict the relevancy $F^i_{pred} = f_{k\_pred}([\text{avg}(\mathbf{F^i_h});I])$. Using the predicted binary classes F^i_{pred} as a mask, we sum pool the fact encoding $\mathbf{F^i_{h'}}$ and compute the final fact encoding representation $\mathbf{F_h} = \sum_i \text{argmax}(F^i_{pred})\mathbf{F^i_{h'}}$. The final fact representation is concatenated with the context encoding, to generate the final hidden representation from the encoder $\mathbf{C_{emb}} = [\mathbf{C_h};\mathbf{F_h}]$, which is passed to the decoder.

5.2 Decoder

We condition the response generation on the policy containing control codes, which enables the model to adapt to the required characteristics. 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

We perform multiple experiments and ablation studies on our proposed system, and use automatic metrics and human judgement for evaluation.

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 backbone 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. All the encoders and the decoder were initialized with the pre-trained backbone model weights, along with parameter sharing for the embedding and token type layers. All models were trained using PyTorch (Paszke et al., 2019), over 2 Nvidia RTX A5000 GPUs using mixed precision (Micikevicius et al., 2018) and learning rate of 2E-5, till the validation loss stopped improving. We utilized 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. As per initial results, the corpus based codes are only input to the encoder to enhance decision making, and are not used as control codes.

6.2 Metrics

We employ both automatic and human evaluation for model comparison. For automatic evaluation, we compare LM 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. Further, in order to measure the faithfulness of the response to the policy, leveraging the annotation models we calculate Intent F1: The average F1 across all intent classes between intent exhibited by the generated text, and the intent of the golden response, and Trait Correlation: The average Pearson's correlation across all Big-5 trait intensities exhibited by the generated response, and the intensities in the golden response. Although the diverse automatic metrics evaluate the models from different perspectives, we also leverage human judgement for evaluating relevance.

6.3 Results and Ablation Study

Leveraging gold knowledge and policy, we report our results and compare with baselines in Table 2. For both WOW and TC, we consider the models using only context and facts as the internal baseline (underlined), and further perform ablation studies to not only gauge the contribution of each policy, but also compare the effectiveness of using Pandora and Essays datasets based control codes as policy. We perform the following ablations: (i) Intent: Using only intent control codes as policy. (ii) C-Traits: Only Corpus based traits in the encoder, no policy. (iii) P / E-Traits: Only Pandora or Essays based personality control codes as policy. (iv) Intent + P / E-Traits: Both Intent and personality control codes as policy. (v) All: Using all control codes as policy, and corpus based traits in the encoder. 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. Further, in order to evaluate the planning capabilities of the models, we compare the F1 score between the predicted policy, and the actual labels on the test data, and report the best performing models by each policy component in Table 3.

6.4 Human Evaluation

We also leveraged human judgement to evaluate our system against our internal baselines, and considered only the BART based models for human evaluation, as it resulted in better BLEU and ROUGE scores. For each dataset, we sampled 100 total examples from both the splits, and asked 2 human evaluators per example to rate each candidate responses by relevance with respect to the conversation context, on a Likert scale from 1 (low relevance) to 5 (high relevance), where the candidates comprised of the response from the baseline and ablation models. 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 sig-

Corpus	Model	Perplexity	BLEU 4	RougeL	BLEURT	Intent F1	Trait Correl.
	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	0.300 / 0.319	0.850 / 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**
WOW	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**
	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	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**
TC	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 2: 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 & golden policy.

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 3: 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.

nificantly 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.

6.5 Observations and Discussion

From the results we observe that leveraging intent and personality trait based control codes as policy, outperform both internal and external baselines,

Context
Agent 1: do you like to party?
Response
<weak_agr><weak_ext><subj><subj_Q> i do not. do you?
<strong_agr><strong_ext><subj><subj_Q> i love to party! do you?

Context
Agent 1: fred missed the penalty, and was dropped from the team. I wonder what he is going through.
Response
<weak_agr><subj> i'm not sure what he's going through right now. i'm sure he is struggling.
<strong_agr><subj> i'm sorry to hear that. i'm sure he's going through some tough times. i hope he's ok.

Context
Agent 1: do you want to go on a hike this weekend?
Agent 2: sure. where are we going?
Agent 1: how about yosemite?
Response
<strong_opn><subj> yosemite is a beautiful place. i would love to hike there.
<strong_opn><subj><obj> 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.

Table 4: Generation examples with different combinations of control codes as policy.

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 (bolded) with p-value < 0.05 and < 0.02 respectively. The baseline result is underlined.

which validates the efficacy of our proposed approach for configurable response generation. Further, we notice that pre-trained BlenderBot results in best perplexity scores, but worse precison/recall metrics, which we attribute to it's low vocabulary size. We also observe that both the Essays and Pandora based codes work well, depending on the scenario. For policy prediction, models incorporating all the control codes seems to perform better, and the presence of personality based features in the context enhances intent prediction. Further, the results indicate that fact selection is a difficult problem, specially for Topical Chat, where the interlocutors have multiple viable options. We further plot the context length wise style adaptation of the generated response in Figure 3, which hints lengthier context facilitates better adaptation to the desired response style.

In Table 4 we showcase a few generated with varied policy configurations against the best performing BART based model trained on WOW. The first 2 examples depict scenarios where varying the agreeableness and extraversion traits results in different response, with the model generating introverted response for weak extraversion, and more empathetic response for strong agreeableness. The third example showcases the model's capability of leveraging external facts. We also observe the capability of the model to adapt to the intent: In the first sample, the model also follows the intent control code and generates a subjective question.

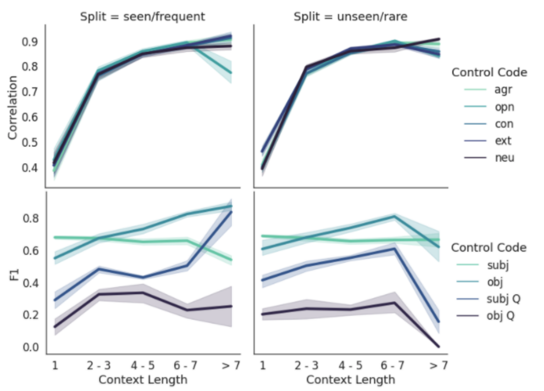

Figure 3: Turn length wise adaptation to the desired response style, collated from all the full version models.

7 Conclusion

In this paper, we experiment with training end-to-end systems, that can not only leverage noisy estimates of Big-5 personality traits, and dialogue intent based control codes as policy for response generation, but also predict the response policy. Our results indicate that the proposed method does generate personality faithful responses, which adheres to the required discourse intents. We establish the efficacy of the system by performing ablation study and comparing automatic metrics against strong internal and external baselines. Further human evaluation demonstrates the benefit of our proposed system in adapting to the required policy.

References

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

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.

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*.

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*.

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.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. In *International Conference on Learning Representations*.

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.

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*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037.

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.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

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, in order to maintain parity between both 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 $\boldsymbol{m}$ non-overlapping samples by randomly selecting and concatenating $\boldsymbol{k}$ sentences, where $\boldsymbol{k}$ was randomly selected to vary between 2 and 30. The target intensities for each of the Big-5 traits were kept same for the $\boldsymbol{m}$ 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 illustrates additional generated samples from both the Wizard of Wikipedia, and Topical Chat test sets. Table 9 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. For each conversation corpus, and personality dataset combination, Table 10 lists the percentage distribution of strong and weak categories (seperated by '/') for each Big-5 trait, by each split of the dataset.

Context
<agent_1>hm, it was so long ago i don't remember now. i caught a big red fish, well big for me. i reeled it in myself
<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.

Best Response
that's true. some hobbyists even make handmade tackle themselves, using plastic lures and artificial flies. that's pretty cool!

Baseline Response
that's a shame! i'm glad you didn't have to use plastic lures or artificial flies to catch the red fish.

Golden Response
yes, it can take quite awhile to get out to the fishing grounds. my goal is to someday make my own tackle! i think i'd like making artificial flies.

Golden Traits: Strong Agr, Strong Opn, Strong Con, Strong Ext, Weak Neu
Golden Intents: Subj, Obj

Context
<agent_1>how's it going, do you follow football? depending on the country it could refer to rugby or soccer

Best Response
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?

Baseline Response
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

Golden Response
love nfl football!

Golden Traits: Strong Agr, Strong Ext, Weak Neu
Golden Intents: Subj

Context
<agent_2>in all honestly though, i do think it's a great story that he was picked 199th in the draft, but he never lost faith in himself and ended-up being one of the greatest quarterbacks of all time!
<agent_1>that is really unbelievable about the draft. the patriots got so lucky. he's had the most super bowl appearances of any player in history!
<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?

Best Response
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?

Baseline Response
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!

Golden Response
i think you're right! you know what else is crazy? six of tom brady's super bowls were decided by 3 or 4 points!

Golden Traits: Strong Agr, Strong Opn, Strong Ext
Golden Intents: Subj, Obj

Table 8: Generation examples from Wizard of Wikipedia and Topical Chat test set.

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 9: 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 control codes.

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