

# PersonalityChat: Conversation Distillation for Personalized Dialog Modeling with Facts and Traits

Ehsan Lotfi, Maxime De Bruyn, Jeska Buhmann, Walter Daelemans

CLiPS Research Center  
University of Antwerp, Belgium  
firstname.lastname@uantwerpen.be

## Abstract

The new wave of Large Language Models (LLM) has offered an efficient tool to curate sizeable conversational datasets. So far studies have mainly focused on task-oriented or generic open-domain dialogs, and have not fully explored the ability of LLMs in following complicated prompts. In this work, we focus on personalization, and employ LLMs to curate a dataset which is difficult and costly to crowd-source: PersonalityChat is a synthetic conversational dataset based upon the popular PersonaChat dataset, but conditioned on both personas and (Big-5) personality traits. Evaluating models fine-tuned on this dataset, we show that the personality trait labels can be used for trait-based personalization of generative dialogue models. We also perform a head-to-head comparison between PersonalityChat and PersonaChat, and show that training on the distilled dataset results in more fluent and coherent dialog agents in the small-model regime.

## 1 Introduction

Training deep learning models and architectures to act as decent dialog agents is a data-hungry task. Historically, big conversational datasets have been either curated from already existing dialog or sudo-dialog resources (e.g. Reddit (Dziri et al., 2019), Ubuntu (Lowe et al., 2015), OpenSubtitles (Lison and Tiedemann, 2016)), or via crowd-sourcing, which is usually the case if specific features are desired (e.g. persona-grounded (Zhang et al., 2018), empathetic (Rashkin et al., 2019)). But these methods are noisy, costly or limited in the scope of feasible grounding features.

Recently, Large Language Models (LLMs) like GPT-3 (Brown et al., 2020), ChatGPT, GPT-4 and OPT (Zhang et al., 2022) provided a cheaper and more flexible alternative, which has been rapidly adopted to augment existing datasets (Chen et al., 2022; Mehri et al., 2022; Sahu et al., 2022) or generating new ones (Rosenbaum et al., 2022; Kim et al.,

<b>Persona:</b> I like to tan in tanning beds. I love spending money. I am high maintenance. I diet a lot. I love listening to Britney Spears.
<b>Personality Traits (ChatGPT):</b> <b>Openness:</b> No clues. <b>Conscientiousness:</b> Probably low, given the statement I love spending money and that they admit to being high maintenance. <b>Extraversion:</b> Probably high, as they enjoy tanning and listening to music, both of which are often done in social settings. <b>Agreeableness:</b> No clues. <b>Neuroticism:</b> Probably high, as they state that they diet a lot and may have insecurities about their appearance.

Table 1: Example of (Big-5) personality traits and explanations generated by ChatGPT when prompted with persona. We then sample the speculated trait labels, and use them along with personas to prompt ChatGPT for personalized dialogs.

2022; Chen et al., 2023). These studies have been mainly focused on task-oriented dialog, or scaling; i.e. exploiting the compelling efficiency and skill of the latest models to generate large-scale quality datasets that can help to mitigate some of the shortcoming of current dialog agents (e.g. common sense knowledge (Kim et al., 2022)). However the potential of these models in assuming sophisticated conversational settings has not been fully explored.

In this study, we exploit this flexibility to distill a corpus which is difficult and expensive to curate from human interactions: a conversational dataset conditioned on both personas and the Big-5 personality traits. To prompt the LLM, we employ persona statements from the popular PersonaChat dataset (Zhang et al., 2018) and enrich them with

relevant labels from the well-known Big-5 personality traits (McCrae and John, 1992) which we also obtain using LLMs. Besides generating the first sizeable dialog corpus curated explicitly with personality traits, this parallel approach allows us to perform a fair comparison between the crowd-sourced PersonaChat and its distilled counterpart.

Our contributions are as follows: 1) We release **PersonalityChat**; a distilled personalized dialog dataset grounded on persona and personality traits. 2) We show that personality trait labels can be used to affect and modify the ‘attitude’ of a dialog agent. 3) We demonstrate that in the small-model regime, training on PersonalityChat results in better performing dialog agents. 4) We also release **PersonaTraits**; a dataset of personality trait speculations, generated by ChatGPT for various personas<sup>1</sup>.

## 2 Related Work

### 2.1 Dialog Distillation from LLMs

Early attempts to generate synthetic dialogues often employed schemes, agendas and templates to simulate data, mostly limited to task-oriented dialog (Schatzmann et al., 2007; Shah et al., 2018; Campagna et al., 2020). Recently, the in-context zero-shot and few-shot learning capabilities of large language models, has made it possible to generate fluent customizable conversations either from scratch, or by providing a few examples. Kim et al. (2022) prompted InstructGPT (Ouyang et al., 2022) with simple narratives based on common sense statements to curate a large social dataset (SODA). Chen et al. (2023) used in-context learning, prompting the OPT (Zhang et al., 2022) model with a combination of topic, facts and dialog examples to generate another social dialog dataset extended to multi-party conversation (PLACES). Lee et al. (2022) proposed a pipeline to augment and update PersonaChat with new personas using GPT-3. More recently, MathDial (Macina et al., 2023) used a semi-synthetic framework to generate dialogs grounded in math reasoning problems by pairing real teachers with LLM-simulated students.

### 2.2 Personalized Dialog Generation

Before neural methods became mainstream, studies like Mairesse and Walker (2007, 2008a,b) and Gill et al. (2012) used parameterised symbolic methods to incorporate the big-5 personality traits in

dialog generation based on lexical features. In one of the first attempts for neural personalised dialog generation, Li et al. (2016) used jointly learned Speaker embeddings to inject personal information into the response. The release of PersonaChat (Zhang et al., 2018) as the first large scale dataset grounded in personal facts, canonised the factual approach to dialog personalization. Curating a sizeable human-written dialog dataset grounded in personality traits is difficult and expensive. Recently Saha et al. (2022) exploited two trait-annotated non-conversational corpora (PANDORA (Gjurković et al., 2021) and Essays (Pennebaker and King, 1999)) to train a classifier which was used to annotate two existing knowledge-based datasets with trait labels. These annotations then were incorporated to train a personalised model.

### 2.3 The Big-5 Personality Traits

Rooted in the Lexical Hypothesis (Galton, 1884) and developed over the last century, the Big Five personality traits (Norman, 1963; Goldberg, 1990), consist of Openness (to experience), Conscientiousness, Extraversion, Agreeableness and Neuroticism. These traits were repeatedly obtained empirically by extracting the main components of a factor analysis over adjective descriptors, and therefore considered as the dimensions explaining the most variance of behaviour among people (Mairesse and Walker, 2008b). Although the Big Five model has limitations (Paunonen and Jackson, 2000), it has strong links to biological bases (Revelle, 2003) and is shown to be a stable personality evaluation metric (Cobb-Clark and Schurer, 2012). Table 11 in the Appendix (from (McCrae and John, 1992)) lists some of the defining adjective and attributes for the five dimensions.

## 3 PersonalityChat

Generating a diverse set of conversations from LLMs is usually done using a varied collection of *scenarios* or settings which provide the model with background information. For example Kim et al. (2022) employed short narratives based on commonsense knowledge, and Chen et al. (2023) used a list of topics and tasks from the FITS dataset (Xu et al., 2022). We base our work on PersonaChat (Zhang et al., 2018), the most popular dataset for dialog personalization which was curated by pairing online crowd-source workers and asking them to converse with each other using their assigned

<sup>1</sup>Datasets will be available at <https://github.com/Elotfi/PersonalityChat>.

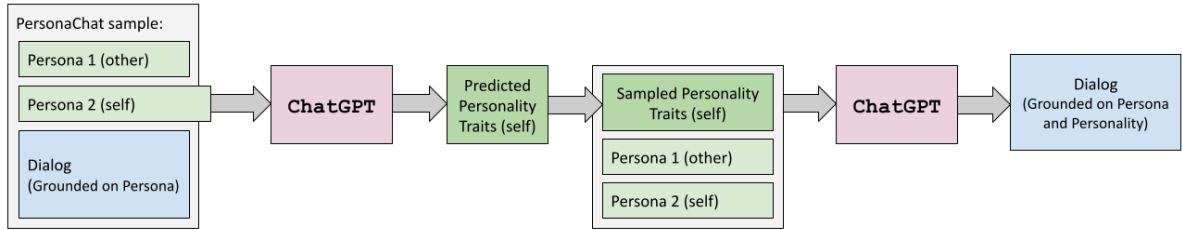


Figure 1: Overall pipeline to curate PersonalityChat. First (left) ChatGPT is used to predict personality traits for PersonaChat personas. Then (right) sampled traits (based on the predictions) are used along with the personas to prompt ChatGPT for dialogs that are grounded on both persona facts and personality traits.

*personas*; a small set of simple personal statements about occupation, interests, hobbies etc. (e.g. Table 1-top). Using PersonaChat allows us to 1) exploit persona statements to acquire (estimated) personality traits, as well as dialog scenarios, 2) perform a head-to-head comparison between crowd-sourcing and distillation, as means of dialog dataset curation.

The overall 2-step curation pipeline is shown in Figure 1: first ChatGPT is employed to predict personality traits for personas from the PersonaChat dataset. Then traits are sampled from these labels and used along with the personas to prompt ChatGPT to generate dialogs grounded on both features. Next we describe these two steps in detail.

### 3.1 Personality Trait Prediction

The PersonaChat dataset (Zhang et al., 2018) comprises of 10,907 dialogs (8939/1000/968 for train/validation/test) between crowd-source workers. Dialog samples come with a pair of *Personas* labeled as ‘self’ and ‘other’, each containing 4-6 personal facts similar to what is shown in Table 1-Top. Not all personas are unique and for example the ‘self’ personas are permuted from 4333/409/414 unique sets.

Personas provide enough background information to prompt LLMs for social dialog and in theory we can just enrich this information with randomly sampled personality traits to generate our prompts. But since persona facts and personality traits are not independent, in order to better align our (trait) sampling with personas, we first use ChatGPT to acquire trait label estimations for each persona, as one of ‘probably low’, ‘probably high’ or ‘no clues’. (The exact prompt, and examples of the predicted labels and supporting explanations can be found in Appendix A) and Table 14 respectively). Figure 2 shows the trait-label distribution, indicating that in 4 out of 5 traits, ‘no clues’ is the majority label. On average personas have 2.3 *indicative* labels (i.e.

‘probably low’ or ‘probably high’)<sup>2</sup>.

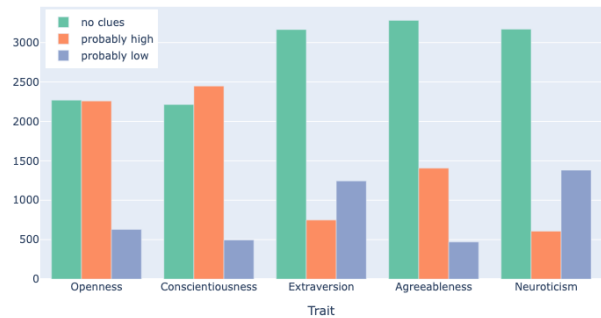


Figure 2: Distribution of the estimated trait labels for PersonaChat personas

Method	Ave. # indicative labels	Ave. Hamming distance	Ave. polar Hamming distance
Human	2.35 ± .14	-	-
ChatGPT	2.27	1.88 ± .06	.13 ± .01
random	3.45	3.42 ± .1	.85 ± .04

Table 2: Comparing human annotations for persona-based traits with ChatGPT and random labels on 100 samples. Distances are measured from the human annotations. Polar Hamming distance only considers polar or extreme disagreement; i.e. high vs. low.

Predicting trait labels from a few facts is inherently noisy; in most cases persona statements do not *strongly* imply or dictate a specific trait, which means that the actual label is a broad distribution. Since we only use these labels for sampling to prompt the LLM (as an alternative to random labels), what is relevant to us is to make sure that: 1) the labels are in fact closer to human annotations than random sampling, 2) they have relatively low polar disagreements (‘high’ vs. ‘low’ for a trait) with the human annotation. To do so, we

<sup>2</sup>We release the acquired dataset as **PersonaTraits**, including 5156 samples of persona, 3-way labels and explanations.

ask two expert annotators to label a set of 100 personas with the same categories ('probably low', 'probably high' or 'no clues'). Table 2 shows how these annotations compare with ChatGPT and random labels. The difference is measured using the Hamming distance (for five traits), and the polar distance only considers extreme disagreement. As can be seen, ChatGPT labels are clearly closer to human annotations, and while the exact agreement is not very high, the polar Hamming distance is quite low (.13 on average). In the next step, we use personas and their acquired trait labels to prompt ChatGPT for social conversations.

## 3.2 Personalized Dialog Generation

### 3.2.1 Sampling Trait Labels

Although we can use labels as they are, we choose to sample based on them to control the complexity, as some personas have 4 or 5 indicative personality labels, which can impose *too many* constraints for a relatively short dialog. In addition, since we are using the same model (ChatGPT) for both steps, we speculate that introducing some randomness would smooth potential lexical links and biases between personas and their estimated traits, and encourage the model to explore more sophisticated ways to reflect the traits in utterances, rather than relying solely on the relevant personal facts. Therefore, for each persona with  $k$  indicative trait-labels (i.e. 'high' or 'low'), we sample  $n$  ( $0 < n < 4$ ) indicative labels, following these rules:

- If  $n \leq k$ , we randomly pick  $n$  labels.
- Otherwise, we compensate the difference by randomly switching  $n - k$  non-indicative labels (i.e. 'no clues') to an indicative one.

### 3.2.2 Prompting

Nest step is prompting ChatGPT with personas and personality traits for grounded dialogs. This can be done with 'single agent' –where one prompt is made with information for both speakers and the model is asked to generate a conversation between them in one pass–, or 'double agent' where each speaker gets their own API (prompted with corresponding information) and the conversation is generated one utterance at a time going back and forth between the agents with updated contexts. In theory the double agent approach sounds more faithful to the real dynamics of a conversation<sup>3</sup> but according to our trials, it often triggers the

<sup>3</sup>As well as the crowd-sourcing method in PersonaChat

'assistant' side of ChatGPT (resulting in long and 'helpful' utterances!) which is hard to mitigate with prompts. Moreover, it is not easy to anticipate or control the length of conversation as it might 'end' before the generation loop is over. The single agent approach on the other hand is simpler and cheaper (on average consuming almost 1/10 tokens) but it is prone to persona 'leakage'; i.e. one speaker referring to the other speaker's information. Another challenge (in both generation settings) is avoiding *explicit* grounding; i.e. cases where the speaker talks about their personality traits in a self-aware and superficial manner (e.g. "I am a very conscientious person."), instead of implicitly reflecting them in responses. We realize that both these issues, can be considerably mitigated by prompt engineering, but not totally eradicated.

After many trials, we opt for the single agent method, plus a lexical post processing step to remove (and regenerate) the most 'leaky' and 'explicit' cases. To control the complexity, we only provide personality traits for the second speaker, corresponding to the 'self' persona (or Person2 in our terminology). We generate a dialog for each PersonaChat sample (i.e. persona pair), picking the temperature uniformly from the (.25, 1.25) interval and limiting the number of generated tokens to 450. The exact prompt can be found in Appendix A.

## 3.3 Dataset Statistics

### 3.3.1 Lexical Features

Table 3 compares various lexical properties of the resulted dataset (PersonalityChat) with those of PersonaChat. **MTLD** is the Measure of Textual Lexical Diversity (McCarthy and Jarvis, 2010), calculated as the mean length of sequential word strings in a text that maintain a given TTR value<sup>4</sup>, and can be interpreted as an indication for local diversity. Based on the numbers, PersonalityChat has a smaller vocabulary<sup>5</sup> but a higher MTLD score, which points to a discrepancy in how the inter- and intra-dialog diversity compare between the two datasets. To see how verbatim the persona facts have been employed in dialogs, we measure the average maximal common n-gram length (**MCN**) between persona facts and utterances or whole di-

<sup>4</sup>We use the same TTR value (.72) as employed by the authors (McCarthy and Jarvis, 2010). For the rationale see the paper.

<sup>5</sup>The fact that PersonaChat is not normalized and contains erroneous and non-standard spellings contributes to this difference.

	PersonaChat	PersonalityChat
Dialogs	10,907	10,907
Turns (avg. dialog)	14.8	17.3
Tokens (avg. turn)	11.7	13.7
MTLD	50.3	69.1
MCN (dialog)	1.76	3.38
MCN (turn)	1.23	2.0
Vocab size (train)	18,669	11,104

Table 3: Lexical comparison between PersonaChat (Zhang et al., 2018) and our distilled PersonalityChat dataset. MTLD stands for Measure of Textual Lexical Diversity (McCarthy and Jarvis, 2010) and MCN is the average maximal common n-gram between persona and dialog (or turn), as described in 3.3.

dialogs<sup>6</sup>, which results in a higher value for PersonalityChat, specially over the whole conversation.

The dialog example in Table 15 demonstrates some aspects of the lexical and stylistic differences: The language usage in the crowd-sourced example (PersonaChat) is more creative, specially in the way that persona facts are being incorporated. This however comes with a significantly higher variance, and while synthesized dialogs almost always follow a safe and streamlined style of fact-grounded personal conversation, the crowd-sourced samples vary between exemplary ones like Table 15, and low quality cases where persona facts are just exchanged with little or no conversational context.

### 3.3.2 Curation Cost

Table 4 shows the generation cost for both steps, i.e. personality trait prediction, and dialog generation, which overall amounts to 15\$ and 70hrs. This excludes trials and corrective re-generations which in our case add an extra 5 %.

	Step 1		Step 2		Overall	
	ps	all	ps	all	ps	all
Time	6s	8.6h	20s	61h	23s	70h
Cost	.05¢	2.6\$	.1¢	12\$	.13¢	15\$

Table 4: Curation cost for both steps in our method; i.e. personality trait prediction (step 1) and dialog generation (step 2). ‘ps’ is the ‘per sample’ value, which does not add up to the last column since each persona is used in almost two dialogs.

<sup>6</sup>For example MCN=3 for dialog means that on average, the longest n-gram shared between individual persona facts and the corresponding dialog has a length of 3 (tokens).

## 4 Training with PersonalityChat

In this section we perform experiments to answer the following questions:

- **RQ1:** Can we use personality traits to control the conversational behavior of a model?
- **RQ2:** How do the models trained on PersonaChat and PersonalityChat compare in performance? (only using persona)

### 4.1 RQ1: Controlling Personality Traits

To answer the first question, we fine-tune a BART model<sup>7</sup> (Lewis et al., 2020) on our dataset, using both personas and personality traits in the context. Following the common practice<sup>8</sup>, we compose the inputs by concatenating different segments as <traits> {trait labels} <persona> {persona facts} {history}, where history includes previous turns marked by <user> and <agent>, and <traits>, <persona>, <user> and <agent> are special tokens added to the vocabulary. We do the fine-tuning using HuggingFace’s transformers library (Wolf et al., 2020) for 8 epochs, while the validation loss declines (More details in Appendix B).

To see if the trained models are capable of generating trait-adapted responses, we perform two sets of evaluations:

#### 4.1.1 Turn-wise Automatic Evaluation

For this part, we use the whole validation set (~8k samples) and generate responses using single trait labels, or no labels (i.e.  $5 * 2 + 1 = 11$  responses for each sample, the last of which we consider as ‘neutral’). Then we compare a number of indicative lexical/semantic features between the inferred utterances, including average sentiment score, average number of tokens, question marks and exclamation marks per utterance. For sentiment we employ a BERT-base model finetuned on the 5-way review classification and use the class probabilities to calculate an aggregated score<sup>9</sup>. The result (Figure 3) shows a consistent pattern in which utterances generated with higher openness, extraversion, agreeableness and conscientiousness, or lower neuroticism are more expressive (a and d), positive

<sup>7</sup>The large version accessible via HuggingFace’s transformers as facebook/bart-large

<sup>8</sup>Using specific tokens and phrases to control language generation is inspired by similar works, specially CTRL (Keskar et al., 2019).

<sup>9</sup>More specifically we calculate  $\sum_{i=0}^4 s[i] * p_i$  where  $s = [-1, -.5, 0, .5, 1]$ .

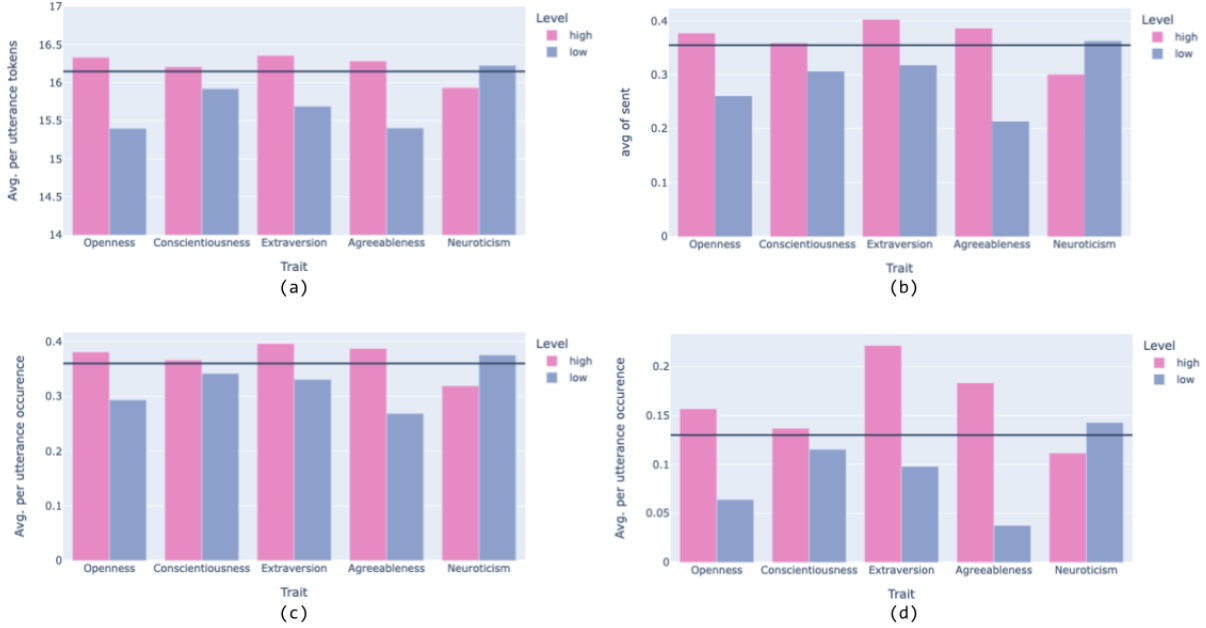


Figure 3: Comparing single-trait utterances based on simple lexical/semantic features: a) number of tokens, b) sentiment score, c) number of question marks, d) number of exclamation marks. The horizontal black line indicates the value in utterances generated without specifying any traits. Values are average over utterances.

Model	Openness			Conscientiousness			Extraversion			Agreeableness			Neuroticism		
	high	tie	low	high	tie	low	high	tie	low	high	tie	low	high	tie	low
Ave. vote	66%	18%	16%	36%	40%	24%	64%	20%	16%	72%	14%	14%	42%	46%	14%
Cohen's $\kappa$	0.56			0.29			0.46			0.49			0.25		

Table 5: Pairwise per-dialog comparison by human annotators between dialogs generated using opposing trait labels (e.g. high and low openness). The second row shows Cohen's kappa between annotators.

(b) and engaging (c), with the difference being relatively higher for openness, extraversion and agreeableness. It also shows that overall, the 'low' label is more effective in steering the response away from the neutral behavior (i.e. the no-trait baseline marked with a black line in each graph).

To have a more semantic view, we identify the frequently 'preferred' and 'avoided' n-grams for each trait-label by comparing n-gram frequencies in the trait and neutral utterances. The result (Table 16 in the Appendix) shows that the statistically indicative n-grams for each trait-label combination<sup>10</sup> generally agree with common semantic expectations for that trait-label (cf. Table 11), although the diversity and richness varies quite noticeably between different traits and labels. In particular, the 'low neuroticism' corpora does not offer any particularly important n-grams.

<sup>10</sup>This is measured by  $abs(x-y)/y$  conditioned on  $abs(x-y) > 20$  where  $x$  and  $y$  are the occurrence frequency in the trait and neutral corpora, respectively.

#### 4.1.2 Pairwise Per-dialog Human Evaluation

To see if the observed statistical patterns lead to discernible semantic differences during interaction, we conduct a human evaluation. Following Smith et al. (2022) who argue for the advantage of per-dialog comparison (vs. per-turn) in situations where differences between models only emerge after several conversation turns, we simulate 100 full conversations with random lengths between two models, one as user (trained on the 'odd' utterances without traits), and one as agent (trained on the 'even' utterances with traits), using persona pairs from the validation set. To mitigate the persona-trait bias, we filter the persona pool, keeping only personas with maximum 2 indicative trait labels (i.e. labels other than 'no clues'. See 3.1). For each sampled persona pair, we generate a pair of conversations for every trait, one with 'high' and one with 'low' label. This results in 500 dialog pairs, generated with single opposing trait labels.

To evaluate the outcome, we ask two annota-

\Eval	PChat					PTChat					Both				
	PPL	F1	TTR	MCN	H	PPL	F1	TTR	MCN	H	PPL	F1	TTR	MCN	H
PChat	9.88	.18	.71	4.75	1.3	21.5	.21	.76	3.58	1.2	15.7	.19	.73	4.20	1.3
PTChat	56.3	.17	.92	4.20	2.0	3.52	.30	.91	3.44	2.2	10.8	.24	.92	3.81	2.1
Both	9.69	.19	.80	4.15	1.8	3.40	.31	.92	3.14	2.3	5.21	.25	.86	3.65	2.1

Table 6: Performance metrics for various train (rows)-evaluation (columns) settings, using **T5-small**. PPL and TTR are perplexity and type/token ratio. MCN is the average length of the maximal n-gram shared between response and persona facts, and H is overall score (from 1 to 3) assigned by human annotators.

Domain	Naturalness			Coherence			Consistency			Specificity			Overall		
	PT	tie	P	PT	tie	P	PT	tie	P	PT	tie	P	PT	tie	P
PChat	70%	26%	4%	49%	40%	11%	25%	64%	11%	12%	67%	21%	74%	15%	11%
PTChat	66%	34%	0%	64%	36%	0%	29%	62%	10%	13%	70%	17%	81%	17%	2%
Ave.	68%	30%	2%	56%	38%	6%	26%	64%	10%	12%	69%	19%	77%	16%	7%

Table 7: Head-to-head quality comparison by human annotators between the output of **T5-small** models fine-tuned on PersonaChat (P), and PersonalityChat (PT). Rows indicate the test data domain.

\Eval	PChat					PTChat					Both				
	PPL	F1	TTR	MCN	H	PPL	F1	TTR	MCN	H	PPL	F1	TTR	MCN	H
PChat	7.56	.21	.85	3.82	2.3	13.9	.25	.88	2.50	2.1	10.9	.23	.87	3.16	2.2
PTChat	52.0	.18	.94	3.36	2.7	2.60	.36	.94	2.95	2.8	8.81	.27	.94	3.15	2.8
Both	7.52	.21	.88	3.40	2.4	2.55	.38	.94	2.79	2.7	3.96	.30	.91	3.10	2.6

Table 8: Performance metrics for various train (rows)-evaluation (columns) settings, using **T5-base**.

Domain	Naturalness			Coherence			Consistency			Specificity			Overall		
	PT	tie	P	PT	tie	P	PT	tie	P	PT	tie	P	PT	tie	P
PChat	23%	73%	4%	40%	49%	11%	19%	79%	2%	11%	75%	13%	60%	15%	25%
PTChat	20%	80%	0%	64%	30%	6%	21%	72%	7%	19%	79%	2%	79%	10%	11%
Ave.	21%	77%	2%	52%	40%	9%	20%	76%	4%	15%	77%	8%	69%	13%	18%

Table 9: Head-to-head quality comparison by human annotators between the output of **T5-base** models fine-tuned on PersonaChat (P) and PersonalityChat (PT). Rows indicate the test data domain.

tors—one psychology student familiar with the topic, and one in-house annotator provided with generic explanatory material on the Big-5 traits—to indicate in which dialog of a pair, the agent shows a *higher* value for that trait, with a possibility for a tie. Table 5 shows the average vote, along with annotator’s agreement (Cohen’s kappa) for each trait, based on which we can conclude that the model is capable of noticeably reflecting Openness, Extraversion and Agreeableness in its output, while Conscientiousness and Neuroticism are not incorporated enough. This is consistent with the automatic results (Figure 3). Table 17 (in the Appendix) shows 5 pairs from our evaluation set, cherry-picked to demonstrate how the model’s outcome is affected by trait labels.

## 4.2 RQ2: PersonaChat vs. PersonalityChat

To investigate the second question, we fine-tune sequence-to-sequence pretrained transformers on these data (without the trait labels), and compare the resulting models using automatic metrics as well as human evaluation. Since we are interested in the low-parameter regime, we first select T5-small<sup>11</sup> which is one of the smallest pretrained encoder-decoders with around 77M parameters. For a fair comparison, we train on the odd utterances (first speaker), since their responses are not grounded in personality traits in our dataset. For brevity, we will refer to PersonaChat and PersonalityChat as **PChat** and **PTChat** respectively, in the

<sup>11</sup>More precisely, we use T5-v1.1-small which is an improved version.

rest of this section.

To evaluate and compare the models, we infer responses from them for the PChat and/or PTChat validation set, using a beam-search of size 5. For automatic metrics, we employ **PPL** (validation perplexity), **F1** (F1 measure between generated and gold response), **TTR** (type/token ratio) and **MCN** (described in 3.3.1). We also ask in-house human annotators to score a 200-sample subset of generated responses (100 for each domain) from 1 to 3, in terms of overall quality<sup>12</sup> (**H** column).

Table 6, summarizes the results for 9 different settings, depending on the train (rows) and evaluation (columns) domains, which can be PChat, PTChat or both. As can be seen, even with T5-small the in-domain perplexity of the PTChat model drops to the very low value of 3.52 (as opposed to 9.88 for PChat model). In terms of generation quality, training on PTChat (second row) generally results in better metrics, even on the out-of-domain PChat data. In particular, the PTChat model (second row) generates significantly less repetitive responses than the PChat model (higher TTR), and acquires a lower MCN score, which is interesting considering the higher MCN value in the PTChat dataset (Table 3). Table 6 also shows that training on both datasets (last row) improves the PChat performance more than PTChat (PChat column: row 3 vs. row 1, compared to PTChat column: row 3 vs. row 2).

To have a more detailed result, we perform head-to-head comparison between the outputs of these models, from the same 200 samples: We ask our annotators to pick the ‘better’ response in terms of Naturalness, Coherence, Consistency, Specificity and Overall quality, with a possibility for a tie<sup>13</sup>. Table 7 shows that the PT model (trained on PersonaChat) beats the P model (trained on PersonaChat) on both domains in the overall preference, as well as specific features, except for Specificity. In particular, PT generates significantly more natural and coherent responses than P. Table 12 shows this advantage still holds in some categories over the PPT model (trained on both datasets), although with a narrower margin.

To see whether these results generalize to bigger models, we repeat the experiments using T5-base (~220M parameters). The results (Tables 8, 9 and 13) show that while PChat benefits more from the

<sup>12</sup>3 being a response that is fluent, coherent, consistent and specific. More details in Appendix C.

<sup>13</sup>More details in Appendix C.

increase in model size, PTChat still holds its overall advantage.

## 5 Conclusion

In this work we exploited the highly customizable generation skills of ChatGPT to build a corpus which is difficult and expensive to curate from human interaction; i.e. a sizeable dialog dataset grounded in both personal facts and (Big-five) personality traits. Starting from the PersonaChat dataset, we enriched the persona statements with personality trait speculations and used them to prompt ChatGPT for personal dialog. We explored the qualities of this dataset in two parts: 1) trait modification, which showed that the overall ‘attitude’ of a dialog agent trained on our dataset can be modified using trait labels, specially for Openness, Extraversion and Agreeableness. 2) comparison with PersonaChat (using only persona facts) made possible by the parallel one-to-one curation scheme, which showed our distilled dataset can be used to train smaller dialog models with better performance, specially in terms of naturalness and coherence.

We also faced and discussed some of the potential shortcomings of our approach. Most importantly, the ‘easy-training’ advantage of a distilled dataset comes with a less diverse and more predictable language distribution which can make a dialog agent sound ‘boring’ on the long run. Combining the natural and synthetic datasets can be a solution but (at least in our experiments) the distilled distribution seems to mostly overrule the natural one. In trait modification, we observed interesting results in the way trait labels can affect and steer model’s conversational behavior but there is a lot of room for improvement (e.g. Table 5). One idea for example might be making use of the explanatory PersonaTraits dataset, which we only released as a side product in this study.

## Limitations

The first limitation of our work regards various biases that are incorporated in the curated data as the result of our choices. Most importantly ChatGPT introduces its (regularised) preferences, specially in dialog generation. Employing a non-proprietary model would have given us more control over the generation process but this comes with significantly less flexibility, and a necessary post-processing step to check for outliers and harmful content. On a



more subtle level, using PersonaChat personas to seed personality traits is not ideal since their generally expressive content and cold/neutral tone, bears implications regarding personality traits which affects the trait distribution.

Our evaluation process, although standard and relatively comprehensive, has limitations which are mostly common in dialog assessment. The automatic metrics are calculated over responses generated from validation or test context which does not necessarily generalise to the real world interaction with an agent. More particular to our work, is the way we evaluated trait modification by using only one trait label at a time, which although makes calculations easier and more reliable, does not necessarily reflect the model’s behavior when prompted with multiple trait labels. The potential persona-personality correlation is another influencing parameter in this step, which although we tried to mitigate by using more neutral persona sets, is not completely resolved.

Finally and as mentioned in the previous section, the trait incorporation in generated dialogs is still basic and rather superficial, which only leaves more room for future studies and improvements.

## Acknowledgements

This research received funding from the Flemish Government under the “Onderzoeksprogramma Artificial Intelligence (AI) Vlaanderen” programme.

## References

- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Giovanni Campagna, Agata Foryciarz, Mehrad Moradshahi, and Monica Lam. 2020. [Zero-shot transfer learning with synthesized data for multi-domain dialogue state tracking](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 122–132, Online. Association for Computational Linguistics.
- Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, and Dilek Hakkani-Tür. 2023. [Places: Prompting language models for social conversation synthesis](#). In *EACL 2023*.
- Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Andy Rosenbaum, Seokhwan Kim, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2022. [Weakly supervised data augmentation through prompting for dialogue understanding](#). In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*.
- Deborah A. Cobb-Clark and Stefanie Schurer. 2012. [The stability of big-five personality traits](#). *Economics Letters*, 115(1):11–15.
- Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. [Augmenting neural response generation with context-aware topical attention](#). In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 18–31, Florence, Italy. Association for Computational Linguistics.
- F Galton. 1884. Measurement of character. fortnightly review, 36, 179-185. *Galton17936Fortnightly Review1884*.
- Alastair Gill, Carsten Brockmann, and Jon Oberlander. 2012. [Perceptions of alignment and personality in generated dialogue](#). In *INLG 2012 Proceedings of the Seventh International Natural Language Generation Conference*, pages 40–48, Utica, IL. Association for Computational Linguistics.
- 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.
- Lewis R. Goldberg. 1990. An alternative "description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59 6:1216–29.
- Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. [Ctrl: A conditional transformer language model for controllable generation](#). *ArXiv*, abs/1909.05858.
- Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2022. [SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization](#). *arXiv e-prints*, page arXiv:2212.10465.
- Young-Jun Lee, Chae-Gyun Lim, Yunsu Choi, Ji-Hui Lm, and Ho-Jin Choi. 2022. [PERSONACHATGEN: Generating personalized dialogues using GPT-3](#). In *Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge*, pages 29–48, Gyeongju, Republic of Korea. Association for Computational Linguistics.

- 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.
- Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. [A persona-based neural conversation model](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.
- Pierre Lison and Jörg Tiedemann. 2016. [OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).
- Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. [The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems](#). In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.
- Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. [Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems](#).
- 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, Prague, Czech Republic. Association for Computational Linguistics.
- François Mairesse and Marilyn Walker. 2008a. A personality-based framework for utterance generation in dialogue applications. pages 80–87.
- François Mairesse and Marilyn A. Walker. 2008b. Can conversational agents express big five personality traits through language ? : Evaluating a psychologically-informed language generator.
- Philip M. McCarthy and Scott Jarvis. 2010. Mtd, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42:381–392.
- Robert R. McCrae and Oliver P. John. 1992. [An introduction to the five-factor model and its applications](#). *Journal of Personality*, 60(2):175–215.
- Shikib Mehri, Yasemin Altun, and Maxine Eskenazi. 2022. [LAD: Language models as data for zero-shot dialog](#). In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 595–604, Edinburgh, UK. Association for Computational Linguistics.
- Warren T. Norman. 1963. Toward an adequate taxonomy of personality attributes: replicated factors structure in peer nomination personality ratings. *Journal of abnormal and social psychology*, 66:574–83.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#).
- S. V. Paunonen and Douglas N. Jackson. 2000. What is beyond the big five? plenty! *Journal of personality*, 68 5:821–35.
- James W. Pennebaker and Lynda A. King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77 6:1296–312.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. [Towards empathetic open-domain conversation models: A new benchmark and dataset](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.
- William Revelle. 2003. [Personality processes](#). *Annual Review of Psychology*, 46:295–328.
- Andy Rosenbaum, Saleh Soltan, Wael Hamza, Yannick Versley, and Markus Boese. 2022. [LINGUIST: Language model instruction tuning to generate annotated utterances for intent classification and slot tagging](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 218–241, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Sougata Saha, Souvik Das, and Rohini Srihari. 2022. [Stylistic response generation by controlling personality traits and intent](#). In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 197–211, Dublin, Ireland. Association for Computational Linguistics.
- Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. [Data augmentation for intent classification with off-the-shelf large language models](#). In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 47–57, Dublin, Ireland. Association for Computational Linguistics.

- Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. [Agenda-based user simulation for bootstrapping a POMDP dialogue system](#). In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 149–152, Rochester, New York. Association for Computational Linguistics.
- Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gokhan Tür. 2018. [Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51, New Orleans - Louisiana. Association for Computational Linguistics.
- Eric Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022. [Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents](#). In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 77–97, Dublin, Ireland. 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. [Trans-formers: 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.
- Jing Xu, Megan Ung, Mojtaba Komeili, Kushal Arora, Y-Lan Boureau, and Jason Weston. 2022. [Learning new skills after deployment: Improving open-domain internet-driven dialogue with human feedback](#).
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. [Personalizing dialogue agents: I have a dog, do you have pets too?](#) In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. [Opt: Open pre-trained transformer language models](#).

## A Appendix: Prompts

For the first step (trait speculation), we prompt ChatGPT with: These are some personal statements about/by <A>:{persona} Can you give a rough speculation on how <A> might score in the Big Five personality traits? Try to summarize the response for each category as one of 'probably low', 'probably high' or 'no clues'. Mentioning 'rough speculation' proves to be crucial to get labels since otherwise the model tends to refuse an answer, pointing out the impossibility to label people based on limited data.

For the second step (dialog generation), we build the prompts from the following template: Generate a conversation between Person1 and Person2 with the following personas:\n Person1: {persona1}.\n Person2: {persona2}. Also bear in mind that Person2 has {sampled\_trait\_labels}. Person2 is NOT self-aware about these traits (so DO NOT directly talk about them) but they can be grasped indirectly from their responses. \n They are chatting online but they don't know each other beforehand and they do NOT see each other's personas.\n Continue the conversation for 12-16 utterances. Keep the utterances natural and short (maximum 2 sentences). The '12-16' constraint is the same as for PersonaChat and sampled\_trait\_labels refers to the verbalized labels, e.g. 'high openness and low extraversion').

## B Appendix: Training Details

We train the models using the trainer method in HuggingFace's transformers library, for 8 epochs with an effective batch size of 128 and a linearly decaying learning rate with a maximum value of 1e-4 for T5 and 2e-5 for BART models. In building the samples, we keep the last 4 turns in the context.

## C Appendix: Human Evaluation (RQ2)

For the human evaluation of generated responses, we presented two in-house annotators with response pairs inferred from compared models using the same input, and asked them to:

- Indicate which response (if any) is more:

**Natural**; i.e. meaningful, fluent and non-repetitive utterance with correct grammar

**Coherent**; i.e. the utterance is a valid continuation of the previous context (disregarding consistency)

**Consistent**; i.e. the utterance is not contradicting any previously stated facts including agent's persona

**Specificity**; i.e. the utterance is not a generic and bland phrase and makes use (if possible) of the relevant facts.

**Overall**; the utterance is preferred in general.

- Score the overall quality of each response with 1,2 or 3, where:
  - 3** means a response that is natural, coherent, consistent and specific.
  - 2** means a response that lacks in one of the four features above.
  - 1** means a response that lacks in two or more of the four features.

Table 10 shows the Cohen's kappa value as a measure of annotators' agreement for different tests reported in Section 4.2.

## **D Appendix: Additional Tables**

Experiment	Nat.	Coh.	Con.	Spc.	Ovr.	PT score	P/PPT score
T5-small, PT vs. P (Table 7)	.73	.66	.82	.68	.77	.67	.71
T5-small, PT vs. PPT (Table 12)	.84	.77	.83	.70	.69	.65	.56
T5-base, PT vs. P (Table 9)	.65	.84	.78	.7	.74	.71	.71
T5-base, PT vs. PPT (Table 13)	.78	.71	.82	.76	.75	.67	.70

Table 10: Cohen’s kappa for different evaluation experiments. P, PT and PPT respectively refer to models trained on PersonaChat, PersonalityChat and both.

Personality Trait	Adjectives	Scales
Openness	Artistic, Curious, Imaginative, Insightful, Original, Wide interests	Fantasy, Aesthetics, Feelings, Actions, Ideas, Values
Conscientiousness	Efficient, Organized, Planful, Reliable, Responsible, Thorough	Competence, Order, Dutifulness, Achievement Striving, Self-Discipline, Deliberation
Extraversion	Active, Assertive, Energetic, Enthusiastic, Outgoing, Talkative	Warmth, Gregariousness, Assertiveness, Activity, Excitement seeking, Positive emotions
Agreeableness	Appreciative, Forgiving, Generous, Kind, Sympathetic, Trusting	Trust, Straightforwardness, Altruism, Compliance, Modesty
Neuroticism	Anxious, Self-pitying, Tense, Touchy, Unstable, Worrying	Anxiety, Hostility, Depression, Self-consciousness, Impulsiveness, Vulnerability

Table 11: Examples of defining adjectives and scales for the big-five traits (from (McCrae and John, 1992))

Domain	Naturalness			Coherence			Consistency			Specificity			Overall		
	PT	tie	PPT	PT	tie	PPT	PT	tie	PPT	PT	tie	PPT	PT	tie	PPT
PChat	41%	53%	6%	36%	40%	24%	11%	74%	15%	13%	68%	19%	49%	17%	34%
PTChat	9%	87%	4%	9%	82%	9%	4%	83%	13%	13%	85%	2%	13%	64%	23%
Ave.	25%	70%	5%	23%	61%	17%	7%	78%	14%	13%	76%	11%	31%	40%	29%

Table 12: Head-to-head quality comparison by human annotators between the output of **T5-small** models fine-tuned on PersonalityChat (PT), and PersonaChat + PersonalityChat (PPT). Rows indicate the test data domain.

Domain	Naturalness			Coherence			Consistency			Specificity			Overall		
	PT	tie	PPT	PT	tie	PPT	PT	tie	PPT	PT	tie	PPT	PT	tie	PPT
PChat	13%	83%	4%	34%	58%	8%	10%	90%	0%	7%	70%	23%	55%	24%	21%
PTChat	4%	96%	0%	15%	81%	4%	0%	98%	2%	6%	92%	2%	21%	66%	13%
Ave.	9%	89%	2%	25%	69%	6%	5%	94%	1%	7%	81%	13%	38%	45%	17%

Table 13: Head-to-head quality comparison by human annotators between the output of **T5-base** models fine-tuned on PersonalityChat (PT), and PersonaChat + PersonalityChat(PPT). Rows indicate the test data domain.

Persona	Speculated Personality Traits (by ChatGPT)
<ul style="list-style-type: none"> <li>- my favorite music genre is death metal.</li> <li>- i am a swim coach at my local pool.</li> <li>- my mother is very traditional while i prefer to be more free spirited.</li> <li>- my family and i are from india.</li> </ul>	<p><b>Openness:</b> Probably high. &lt;A&gt; seems to have a preference for a music genre that is considered unconventional or niche, suggesting an openness to new experiences and unconventional ideas.</p> <p><b>Conscientiousness:</b> No clues.</p> <p><b>Extraversion:</b> No clues.</p> <p><b>Agreeableness:</b> Probably low. The contrast between &lt;A&gt;'s free-spiritedness and their traditional mother suggests a potential conflict in values or beliefs, which could indicate a lower degree of agreeableness or at least a willingness to challenge normative expectations.</p> <p><b>Neuroticism:</b> No clues.</p>
<ul style="list-style-type: none"> <li>- my dream in life is to work from home.</li> <li>- i went to school to be a veterinarian but hated it.</li> <li>- i dye my hair every three months.</li> <li>- i have an internet addiction and spend a lot of time playing roleplaying games.</li> <li>- i faint at the sight of blood.</li> </ul>	<p><b>Openness:</b> Probably low: &lt;A&gt; seems to have a clear goal of working from home and did not enjoy the veterinary field, indicating a preference for stability and practicality over creativity and innovation.</p> <p><b>Conscientiousness:</b> No clues. While &lt;A&gt; did not enjoy their chosen field of study, their desire to work from home suggests a certain level of self-discipline and responsibility.</p> <p><b>Extraversion:</b> Probably low. &lt;A&gt; admits to having an internet addiction and spending a lot of time playing roleplaying games, suggesting a preference for solitary activities rather than social ones.</p> <p><b>Agreeableness:</b> No clues.</p> <p><b>Neuroticism:</b> Probably high. &lt;A&gt; mentions fainting at the sight of blood, indicating a strong emotional response to a particular stimuli. They also mention an internet addiction, which can be a coping mechanism for anxiety.</p>
<ul style="list-style-type: none"> <li>- my boyfriend and i are moving into an apartment together next week.</li> <li>- my favorite colors are red and silver.</li> <li>- i am fluent in english spanish and french.</li> <li>- i am an elementary school teacher.</li> </ul>	<p><b>Openness:</b> Probably high. &lt;A&gt; speaking three languages fluently shows an interest in learning and experiencing different cultures which is an indication of high openness.</p> <p><b>Conscientiousness:</b> Probably high. Being an elementary school teacher requires a great deal of responsibility and organization which indicates high conscientiousness.</p> <p><b>Extraversion:</b> No clues.</p> <p><b>Agreeableness:</b> Probably high. Wanting to move in with a partner and having favorite colors indicates a desire for harmony, which is an indication of high agreeableness.</p> <p><b>Neuroticism:</b> No clues.</p>
<ul style="list-style-type: none"> <li>- my wife spends all my money.</li> <li>- i am 40 years old.</li> <li>- i hate my job.</li> <li>- i work as a car salesman.</li> </ul>	<p><b>Openness:</b> No clues.</p> <p><b>Conscientiousness:</b> Probably low. The statement "my wife spends all my money" implies a lack of control over finances, which could suggest a tendency towards impulsivity and low conscientiousness.</p> <p><b>Extraversion:</b> No clues.</p> <p><b>Agreeableness:</b> No clues.</p> <p><b>Neuroticism:</b> Probably high. The statements "i hate my job" and "my wife spends all my money" suggest a level of dissatisfaction and stress that could indicate high levels of neuroticism.</p>

Table 14: Examples from personality traits speculated by ChatGPT based on personas.

<p><b>Persona 1 (other)</b>  My friends are all skateboarders.  I have a broken arm.  I am a skateboarder.  I am always wearing a hat</p>	<p><b>Persona 2 (self)</b>  I love dogs.  I live in Alabama.  I have three sisters.  I like tacos.</p>
<p><b>PersonaChat:</b>  P1) hello, how are you tonight ?  P2) ok. since Alabama, my home team, won Saturday night.  P1) I am an Ohio state fan. do you skate?  P2) no. too busy with all of my 19 dogs and 3 sisters.  P1) oh wow. that is a lot of dogs.  P2) yes love them almost as my tacos. no i love my dogs more .  P1) my fiance makes a killer fish taco with cilantro lime slaw.  P2) wow. my sisters and i love chicken tacos.  P1) those are good too, i live in Cali along the ocean.  P2) I live in middle of Alabama. do not get to ocean much.  P1) I could not stand to live inland, the ocean runs through my veins.  P2) yes. i want to go more but no one to stay with 19 dogs.  P1) maybe someone will come and dog sit for you.  P2) maybe each sister could take six dogs each and foxy go with me.  P1) there you go. where there is a will, there is a way.</p>	
<p><b>PersonalityChat:</b>  (sampled) Trait labels for Persona 2: low conscientiousness</p> <p>P1) Hey there! How’s it going?  P2) Pretty good, just hanging out with my dog. You?  P1) Not too bad, just recovering from a broken arm.  P2) Ouch, that sounds painful. How did it happen?  P1) Skateboarding accident. But it’s all good, just have to take it easy for a bit.  P2) Oh cool, I’ve always wanted to try skateboarding but I’m too clumsy. Do you like any other sports?  P1) Not really, just skateboarding. What about you?  P2) I’m not really into sports, but I do like to go for walks with my dog.  P1) That sounds nice. What kind of dog do you have?  P2) A golden retriever. She’s the sweetest thing.  P1) Aw, I love dogs. I have a beagle myself.  P2) Nice! Do you have any other pets?  P1) Just the beagle. I’m always wearing a hat though, it’s kind of my thing.  P2) Haha, that’s cool. I like tacos, they’re my thing.  P1) Tacos are pretty great. Do you have a favorite kind?  P2) I like them all, but I really love ones with lots of cheese and guacamole.  P1) Oh man, now I’m getting hungry.</p>	

Table 15: Dialog example from the PersonaChat dataset and our PersonalityChat generated using the same Personas in the prompt.

Openness	High	Preferred	experiences (25,3), open-minded (60,12), explore (36,11), open (90,44), learning (83,52), new (450,299), books (106,73), try new things (42,13), explore new places (29,9), always up for (40,17), important to be open (39,15)
		Avoided	not really into (105,133), not much of a (104,132)
	Low	Preferred	whatever (51,14), stuff (78,22), prefer (295,92), not/n't (2632,1764), not really into (398,133), prefer to stay (35,12), i mostly just (34,14), to be honest i prefer (24,1), i'm not much of a (216, 132), i don't really have a favorite (123,62), i don't think i could handle (52,32)
		Avoided	to try (47,70), i love (493,754), all kinds (37,60), to visit (48,79), your favorite (53,112), that's awesome (67,187), i can appreciate (53,74), thanks for asking (213,369), admired people who (19,62), have you ever been (27,48), i've always wanted (91,184)
Conscientiousness	High	Preferred	organized(37,9), finance(39,19), on top(33,8), stay active(98,66), important to have(101,68)
		Avoided	could (91,130), maybe (65,86)
	Low	Preferred	stuff (83,22), whatever (34,14), yeah (703,468), hang out (46,22), not really into (280,133), just got back from (37, 17), i'm not very good at (73,50)
		Avoided	challenging (18,47), yes (81,195), admired (29,65), dedication (28,52), i understand (49,86), it's important to (183,324), outside of work (20,44), i'm doing well (99,266)
Extraversion	High	Preferred	hey (63,11), awesome (355,192), love animals (42,21), trying new (138,94), <u>i'm doing great</u> (231,10), going to concerts (30,6), always up for (42,17), i love going (92,49)
		Avoided	watch (49,74), <u>i'm doing well</u> (107,266)
	Low	Preferred	not much (274,150), <u>that's nice/interesting</u> (332,228)
		Avoided	<u>that's awesome</u> (87,187), i'm a big fan (39,61)
Agreeableness	High	Preferred	rewarding (83,55), have a great day (125,79), that's great to hear (62,42), sounds like a lot of fun (109,82)
		Avoided	i'm not really into (100,133)
	Low	Preferred	fine (21,1), whatever (87,14), radio (36,6), prefer (527,92), stuff (125, 22), indoors (33,7), responsibility (33,13), bad (113,48), not/n't (2728,1764), sounds boring (81,1), so so (57,13), i just eat whatever (29,1), not really my thing (71,5), i respect people who (87,32), that's too bad (102,39), i don't really care about that (33,0), i don't have much free time (38,14)
		Avoided	sweet (11,48), awesome (46,192), visit (47,99), great (419,827), wow (150,260), sorry (99,154), sounds delicious/amazing (44,106), that's understandable (86,137), i love (501,754), thanks for asking (96,369), i always admired people who (9,62), everyone has their own preferences (53,81), i love spending time with (26,52)
Neuroticism	High	Preferred	anxious (48,2), stressful (41,7), helps me relax (29,8), i wish i (98,46), not really into (206,133), <u>i'm doing alright</u> (279,61), i'm not very good at (82,50), i'm not really into (206,133)
		Avoided	<u>i'm doing well</u> (12,266)
	Low	Preferred	-
		Avoided	-

Table 16: Frequently preferred or avoided n-grams in responses generated with single trait labels. The numbers in front of each entry show the n-gram's frequency in the trait ( $x$ ) and neutral ( $y$ ) response set (7825 samples for each) respectively. Terms are ordered by the  $abs(x - y)/y$  ratio, in increasing n-gram length. Underlined pairs (in the same trait) indicate alternative terms.



Table 17: (cherry-picked) Examples of dialogs generated with single trait labels

Openness	<p><b>Persona:</b>                  I like shopping.                  I often shop for shoes and watches.                  My family does not support my career choices.                  I currently work for a watch dealer.</p>
	<p>High</p> <p>&lt;user&gt; Hi there! What kind of music do you like?                  &lt;agent&gt; Hey! I'm into all kinds of music, but lately I've been really into indie rock. How about you?                  &lt;user&gt; My favorite band is Rascal Flatts. I grew up listening to them all the time.                  &lt;agent&gt; Oh, I've heard of them! I'll have to check them out. So, what do you do for a living?                  &lt;user&gt; I'm a country music singer. I grew up in the industry, so it's always been a big part of my life.                  &lt;agent&gt; That's really cool! I currently work for a watch dealer. It's not the most exciting job, but it pays the bills.                  &lt;user&gt; Yeah, I know what you mean. It's important to have a job that you enjoy. Do you have any hobbies outside of work?                  &lt;agent&gt; Yeah, I love shopping! I often shop for shoes and watches. It's a great way to unwind after a long day.</p> <p>Low</p> <p>&lt;user&gt; Hi there! What kind of music do you like?                  &lt;agent&gt; I don't really listen to music. I prefer shopping for shoes and watches.                  &lt;user&gt; Oh, that's interesting. I'm a country music singer. Do you like country music?                  &lt;agent&gt; No, I don't really like country music. It's too boring for me.                  &lt;user&gt; That's too bad. I grew up in the industry and my favorite band is Rascal Flatts.                  &lt;agent&gt; I don't really listen to country music.                  &lt;user&gt; That's okay. We spent summers at national parks. Have you ever been to any?                  &lt;agent&gt; No, I haven't. I don't really like being outside.</p>
Conscientiousness	<p><b>Persona:</b>                  I am a 45 year old male.                  I now live in the city.                  I take many medications and some of them make me see things.                  I grew up on a farm.                  I am a recovering alcoholic.</p>
	<p>High</p> <p>&lt;user&gt; Hi there! What do you like to do for fun?                  &lt;agent&gt; Hi! I like to read and watch movies. What about you?                  &lt;user&gt; I love cooking and fishing. If I'm not on the water, I'm in the kitchen.                  &lt;agent&gt; That's cool. I grew up on a farm, so I know a thing or two about cooking.                  &lt;user&gt; Oh wow, that's interesting. What kind of dishes did you like to cook?                  &lt;agent&gt; I loved making homemade bread and pies. It was a lot of work, but it was always worth it.                  &lt;user&gt; That sounds delicious. I'm not much of a baker myself, but I love trying new foods. Do you have any favorite foods?                  &lt;agent&gt; I'm a recovering alcoholic, so I try to eat healthy most of the time. I take many medications to help me deal with my addiction.                  &lt;user&gt; I'm sorry to hear that. It takes a lot of strength to overcome addiction. Do you have any tips for staying sober?                  &lt;agent&gt; I try to stay organized and plan my meals ahead of time. It helps me stay on track with my medication schedule.                  &lt;user&gt; That's great advice. It's important to take care of yourself and do what makes you happy.                  &lt;agent&gt; Thanks for the advice. I'll keep that in mind. It was nice chatting with you.</p> <p>Low</p> <p>&lt;user&gt; Hi there! What do you like to do for fun?                  &lt;agent&gt; Hey! I like to watch movies and play video games. What about you?                  &lt;user&gt; I love cooking and fishing. If I'm not on the water, I'm probably in the kitchen.                  &lt;agent&gt; That's cool. I've never been fishing before. Do you catch a lot of fish?                  &lt;user&gt; Yeah, I catch a lot of fish. It's a great way to relax and unwind.                  &lt;agent&gt; I can imagine. I grew up on a farm, so I know what it's like to have a good time.                  &lt;user&gt; Oh wow, that's interesting. What kind of animals did you work with on the farm?                  &lt;agent&gt; Mostly cows and chickens. It was hard work, but I loved it. Do you have any pets?                  &lt;user&gt; No, I don't have any pets right now. But I used to have a dog when I was younger. He was my best friend.                  &lt;agent&gt; Dogs are great. I used to have a dog too, but he passed away a few years ago. It was tough losing him.                  &lt;user&gt; I'm sorry to hear that. Losing a pet can be really tough. Do you have any other hobbies besides playing video games?                  &lt;agent&gt; Not really. I'm a recovering alcoholic, so I don't have a lot of free time. But when I do, I like to watch movies and play video games.                  &lt;user&gt; I'm sorry to hear that. But it's great that you're recovering. It takes a lot of strength to overcome addiction.                  &lt;agent&gt; Yeah, it can be tough sometimes. But I'm trying to stay positive and focus on the good things in life.</p>

Extraversion	<p><b>Persona:</b>  I am a 45 year old male.  I now live in the city.  I take many medications and some of them make me see things.  I grew up on a farm.  I am a recovering alcoholic.</p>
	<p>High</p> <p>&lt;user&gt; Hi there! What do you like to do for fun?  &lt;agent&gt; Hey! I love going to concerts and trying new restaurants. What about you?  &lt;user&gt; I'm a big fan of cooking and fishing. I also love comic books. Spiderman is my favorite.  &lt;agent&gt; Oh, that's cool. I'm not really into comics, but I can appreciate a good story.  &lt;user&gt; Yeah, it's definitely a guilty pleasure of mine. What kind of music do you like?  &lt;agent&gt; I'm pretty open to all kinds of music, but lately I've been really into indie rock. What about you?  &lt;user&gt; I like a little bit of everything, but I'm really into classic rock. Do you have any favorite bands?  &lt;agent&gt; Yeah, I really like Led Zeppelin and Pink Floyd. What about you?  &lt;user&gt; I'm more into indie rock, but I do like some classic rock too. Have you ever been to a concert?  &lt;agent&gt; Yeah, I've been to a few. It's always a great time. Do you have any pets?  &lt;user&gt; No, I don't have any pets at the moment. What about you?  &lt;agent&gt; I used to have a dog, but he passed away a few years ago.</p>
	<p>Low</p> <p>&lt;user&gt; Hi there! What do you like to do for fun?  &lt;agent&gt; I like to read and watch movies. What about you?  &lt;user&gt; I love cooking and fishing. If I'm not on the water, I'm in the kitchen.  &lt;agent&gt; That's cool. I've never been fishing before. Do you catch a lot of fish?  &lt;user&gt; Yeah, I try to catch a few every time. It's a great way to relax and unwind.  &lt;agent&gt; I can imagine. Well, it was nice chatting with you. Have a good day!  &lt;user&gt; You too! Take care."</p>
Agreeableness	<p><b>Persona:</b>  I plan to go to business school next year.  I grew up in Manhattan.  I drive a BMW.  I graduated from Yale in 2011 with a degree in finance.</p>
	<p>High</p> <p>&lt;user&gt; Hi there! How are you doing today?  &lt;agent&gt; I'm doing well, thanks for asking. How about you?  &lt;user&gt; I'm good, thanks. So, what do you like to do in your free time?  &lt;agent&gt; I enjoy driving my BMW around and exploring new places. How about you?  &lt;user&gt; I love playing board games with my friends. It's a great way to unwind after a long day at work.  &lt;agent&gt; That sounds like a lot of fun. I'm not much of a game person myself, but I can appreciate a good game.  &lt;user&gt; Yeah, it's a great way to bond with friends and have some fun. So, what do you do for a living?  &lt;agent&gt; I graduated from Yale in 2011 with a degree in finance and I plan to go to business school next year. What about you?  &lt;user&gt; I'm an elementary school teacher. I love working with kids and helping them learn and grow.  &lt;agent&gt; That's really cool! I've always admired teachers. What subject do you teach?</p>
	<p>Low</p> <p>&lt;user&gt; Hi there! How are you doing today?  &lt;agent&gt; I'm good, thanks. How about you?  &lt;user&gt; I'm doing well, thanks. So, what do you like to do in your free time?  &lt;agent&gt; I like to drive my BMW around and explore new places. What about you?  &lt;user&gt; I love playing board games with my friends. It's a great way to pass the time.  &lt;agent&gt; Board games? That sounds boring to me. I prefer more exciting activities.  &lt;user&gt; Oh, I see. Well, everyone has their own interests. What kind of music do you like?  &lt;agent&gt; I don't really have a favorite genre. I just listen to whatever's on the radio.  &lt;user&gt; That's cool. My favorite band is up and I got to see them in concert once.  &lt;agent&gt; That's nice! I've never been to a concert before.</p>

Neuroticism	<p><b>Persona:</b>  I am a stunt double as my second job.  I was raised in a single parent household.  I only eat kosher.  I never broke a bone in my body ever in my life.</p>
	<p>High</p> <p>&lt;user&gt; Hi there! How are you doing today?  &lt;agent&gt; I'm doing alright, thanks for asking. How about you?  &lt;user&gt; I'm doing pretty well. I just started a new job as a life coach.  &lt;agent&gt; That's cool. I'm a stunt double as my second job.  &lt;user&gt; Oh wow, that sounds like a lot of fun. What kind of stunts do you do?  &lt;agent&gt; Mostly stunt work, but I also do some comedy and acting. It's a lot of fun, but it can be stressful at times.  &lt;user&gt; I can imagine. I'm a cancer survivor, so I know how important it is to take care of yourself and your loved ones.  &lt;agent&gt; I'm so sorry to hear that. I never broke a bone in my body ever in my life, so I know how important it is to take care of yourself.</p>
	<p>Low</p> <p>&lt;user&gt; Hi there! How are you doing today?  &lt;agent&gt; I'm doing well, thanks for asking. How about you?  &lt;user&gt; I'm doing great, thanks. I just started a new job as a life coach.  &lt;agent&gt; Oh, that's interesting. What kind of life coaching do you do?  &lt;user&gt; I work with clients to help them achieve their goals and achieve their dreams.  &lt;agent&gt; That's really cool. I've always admired people who can help others achieve their goals.  &lt;user&gt; Thank you, I really enjoy it. So, what do you like to do in your free time?  &lt;agent&gt; I love to travel and explore new places. I'm always up for an adventure. What about you?</p>