PERSONACHATGEN: Generating Personalized Dialogues using GPT-3

Young-Jun Lee[1] **Chae-Gyun Lim**[1] **Yunsu Choi**[2] **Ji-Hui Lm**[2] **Ho-Jin Choi**[1]
[1]School of Computing, KAIST [2]KT Corporation
{yj2961,rayote,hojinc}@kaist.ac.kr {yunsu.choi,jihui.im}@kt.com

Abstract

Recently, many prior works have made their own agents generate more personalized and engaging responses using PERSONACHAT (Zhang et al., 2018). However, since this dataset is frozen in 2018, the dialogue agents trained on this dataset would not know how to interact with a human who loves "*Wandavision*." One way to alleviate this problem is to create a large-scale dataset. In this work, we introduce the pipeline[1] of creating PERSONACHATGEN, which is comprised of three main components: Creating (1) PROFILEGEN, (2) Persona Set, and (3) PERSONACHATGEN. To encourage GPT-3's generation ability, we also defined a taxonomy of hierarchical persona category derived from social profiling taxonomy (Bilal et al., 2019). To create the speaker consistent persona set, we propose a simple contradiction-based iterative sentence replacement algorithm, named CONL. Moreover, to prevent GPT-3 generating harmful content, we presented two filtering pipelines, one each for PROFILEGEN and PERSONACHATGEN. Through analyzing of PERSONACHATGEN, we showed that GPT-3 can generate personalized dialogue containing diverse persona. Furthermore, we revealed a state-of-the-art Blender 90M trained on our dataset that leads to higher performance.

1 Introduction

Considering users' personal information (e.g., preferences, gender, age, and profession) is an essential capability for chit-chat dialogue agents. Since PERSONACHAT was released in 2018, many studies have attempted to build their own dialogue agents to generate personalized and engaging responses in dialogue. These studies published in ACL Venues usually utilized the PERSONACHAT dataset. However, this dataset was constructed in 2018, so dialogue agents trained on it cannot understand how to interact with users who loved the "*Avengers: End game*" movie, which can be regarded as unseen information. One way to solve this problem is to construct a large-scale dataset that includes more diverse personal information and how to interact with a conversation partner based on them. However, the process of manually creating dataset is time-consuming and costly.

Recently, as an alternate way, many studies have created datasets by leveraging pre-trained language models with designed prompt instructions (Yoo et al., 2021; Baheti et al., 2021; Hartvigsen et al., 2022) due to their enormous ability to produce more human-like text (Clark et al., 2021; Dou et al., 2021). They mainly focused on creating datasets related to NLU tasks, such as text classification, textual similarity, and natural language inference. However, no approach has generated a personalized dialogue dataset using a pre-trained language model, especially GPT-3. Note that our goal is to provide insights that prompting language models can create such datasets, not to release a new dataset generated by a language model.

In this work, we introduce the pipeline of creating PERSONACHATGEN, a small-scale machine-generated dataset of 1,649 dialogues. Motivated by (Mishra et al., 2021) and the collection process of PERSONACHAT, our pipeline consists of three main parts: (1) PROFILEGEN Creation, (2) Persona Set Creation, and (3) PERSONACHATGEN Creation. To obtain high-quality generated results from (1) and (2), we first defined a taxonomy of hierarchical persona category based on the social profiling taxonomy (Bilal et al., 2019). Then, we carefully designed prompts. Since GPT-3 can generate offensive and socially biased text (Baheti et al., 2021; Hartvigsen et al., 2022), we also present filtering steps in our pipeline.

- We introduced a novel pipeline for automatically generating PERSONACHATGEN, that consists of three parts: (1) PROFILEGEN Cre-

[1]Our code is available at https://github.com/passing2961/PersonaChatGen

Proceedings of the 29th International Conference on Computational Linguistic, pages 29–48
October 12–17, 2022.

ation, (2) Persona Set Creation, and (3) PERSONACHATGEN Creation. We can adjust an arbitrary number of dialogue turns, which is a powerful advantage of our proposed pipeline.

- We show that Blender 90M (Roller et al., 2020) trained on PERSONACHATGEN and PERSONACHAT together achieve better performance in both automatic and human evaluation.

- We provide the insight that we can leverage the prompting language model [2] (e.g., GPT-3) to generate personalized dialogues datasets. To the best of our knowledge, this is the first study to automatically generate personalized dialogues using GPT-3.

2 Related Work

Persona Dialogue Generation. Li et al. (2016) encoded persona information into the embedding space. To create more engaging dialogue agents, Zhang et al. (2018) released the PERSONACHAT dataset that was collected from a crowd-sourcing platform (Amazon Mechanical Turk). Madotto et al. (2019) used meta-learning to personalize dialogue agents. Liu et al. (2020) improved the quality of generated responses by incorporating mutual persona perception.

Dataset Generation. Yoo et al. (2021) leveraged GPT-3 to generate datasets for text classification tasks. Schick and Schütze (2021) first released a textual similarity dataset generated using a pre-trained language model (PLM) with instructions. Meng et al. (2022) used a unidirectional PLM to generate a dataset that corresponds to given label information for the zero-shot learning of NLU tasks. Then, they fine-tuned a bidirectional PLM using automatically constructed datasets. However, how to generate persona dialogue datasets remain under-explored in the literature.

3 Preliminaries

In this section, we define *persona* and the main terminologies used in this work.

3.1 Task Formulation

This task aims to generate more consistent responses y conditioned on given dialogue context x and persona set P by maximizing $p(y|x, P) = \prod_t p(y_t|y_1, ..., y_{t-1}, x, P)$, where $P = \{p_i\}_{i=1}^{N}$ and N denotes the number of sentences that the persona set P contains. Since PERSONACHAT (Zhang et al., 2018) is created by two humans who are assigned to each persona set, it contains two persona sets for each dialogue.

3.2 Persona Definition

First, we define a *persona* in this work based on the literature survey. Following the Wikipedia definition [3], a *persona* is simply a fictional character. Li et al. (2016) regarded personas as compositions of identities (background facts or user profile), language behavior, and interaction style. Zhang et al. (2018) defined a persona as a character created by multiple profile sentences. In this work, we define a personas as user profiles. Several works considered each profile sentence (e.g., *I like to play a soccer*) as *personal attribute*, which explicitly represents an identity and characteristics (Welleck et al., 2018; Wu et al., 2019; Wang, 2021). This *personal attribute* is mainly represented in the triple format of (e_1, r, e_2), where e_1, r, and e_2 denote entity 1, relation type, and entity 2, respectively. Herein, we define this relation type as *persona category* and entity 2 as *persona entity*. The *persona entity* is a key-value format. For example, in the *personal attribute* of "*I'm from Boston, MA*", the *persona category* is "location" and the *persona entity* is "(city-state, Boston, MA)".

4 A Taxonomy of Hierarchical Persona Categories

Most previous studies have not explicitly established a taxonomy for the persona category. Welleck et al. (2018) defined various relation types and entity categories (See Appendix F). Furthermore, they presented the hierarchical category for relation types. However, there is significant room to establish more sophisticated categories. We have several reasons for introducing the hierarchical persona category. In the real world, the persona comprises a hierarchical structure. For example, within the "preference" category, there is a preference about "movie" and a further preference about "movie title" or "movie genre." In the practical perspective, we should provide well-designed prompts into GPT-3 to enhance the quality of generated dialogues (Mishra et al., 2021). As we mentioned in

[2] In this work, we use GPT-3 (Brown et al., 2020), but our pipeline could work with any prompting language model, such as OPT (Zhang et al., 2022)

[3] https://en.wikipedia.org/wiki/Persona

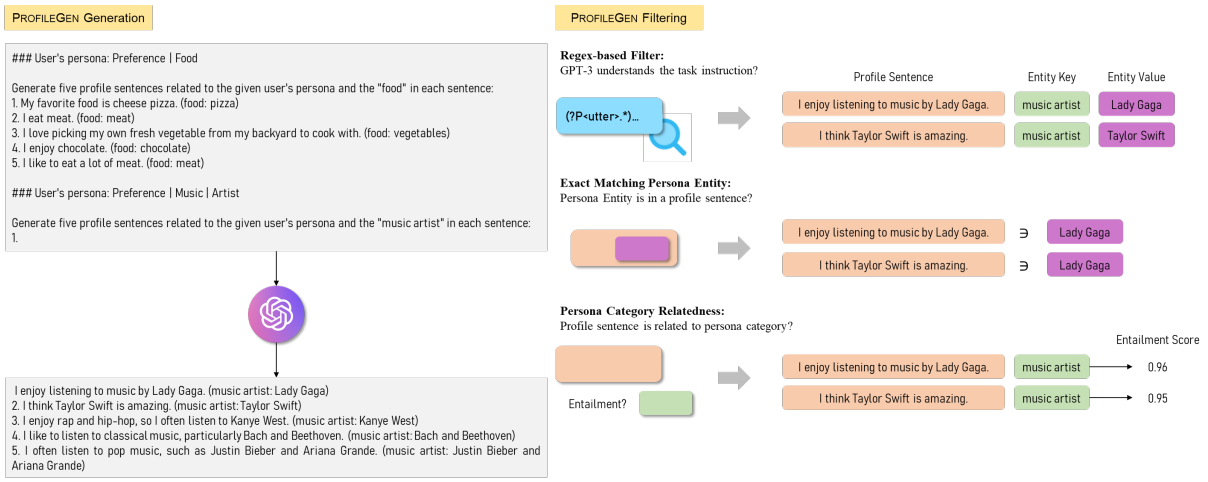

Figure 1: The overall pipeline of PROFILEGEN.

the definition of persona (§3.2), we can regard a persona as a user profile, which can be also viewed as an individual profile. Following (Bilal et al., 2019), we introduce a taxonomy of three main hierarchical persona categories: DEMOGRAPHICS, PSYCHOGRAPHICS, and WELLNESS.

Note that we provide a basic taxonomy of hierarchical persona categories. Appendix E describes these taxonomies in more detail.

5 A Pipeline of Creating PERSONACHATGEN

This section introduces the pipeline of creating PERSONACHATGEN, which consists of three main parts: (i) PROFILEGEN Creation, (ii) Persona Set Creation and (iii) PERSONACHATGEN Creation. To create a consistent persona set, we also propose a simple contradiction-based iterative replacement algorithm, named CONL.

5.1 DECOMPOSITION REFRAMING-based Prompt Engineering

Generating a personalized dialogue dataset from scratch using GPT-3 is challenging for two likely reasons: if a target task itself is inherently difficult or if the task instruction itself is complicated, thus a prompting language model (e.g., GPT-3) cannot achieve higher performance, as reported in Mishra et al. (2021). Furthermore, since the datasets used in GPT-3 pre-training are mainly formal languages (e.g., books and Wikipedia), the generative probability distribution itself learned in GPT-3 will be biased toward formal language. Therefore, we should design prompts to be intuitive and understandable from the GPT-3's perspective.

To make a prompt suitable for creating PERSONACHATGEN, we ponder: "*how was* PERSONACHAT *collected?*" First, they collected 1,155 persona sets. Each persona set P consists of multiple profile sentences (i.e., four or five sentences) and each sentence is written by Turkers. Then, two Turkers chat to get to know one another, where the persona set P is randomly assigned to each Turker.[4] Inspired by this collecting process, we decompose our task into two different sub-tasks, which is similar to the DECOMPOSITION REFRAMING techniques in Mishra et al. (2021). Both creating PROFILEGEN and PERSONACHATGEN parts equally include (1) *generation* describes how GPT-3 generates contents with our designed prompt and (2) *filtering* describes how we remove unreasonable content to enhance the quality of PERSONACHATGEN.

5.2 PROFILEGEN Creation

Here, we describe how we create PROFILEGEN; Figure 1 illustrates the overall process.

5.2.1 Generation

We utilized GPT-3 to create the persona set consisting of multiple profile sentences. In PERSONACHAT, when collecting several profile sentences, the researchers did not explicitly instruct the Turkers to generate sentences corresponding to given *persona categories*. However, in this study, the *persona category* should be explicitly indicated so that GPT-3 understands the given task well. Therefore, we carefully designed the tax-

[4] We omitted the revised personas process, which was originally described in PERSONACHAT (Zhang et al., 2018).

onomy in Section 4. Table 12 and 13 show the prompt template for PROFILEGEN and an example of the constructed prompt with generated profile sentences, respectively.

5.2.2 Filtering

To obtain high quality results, we present a filtering pipeline for PROFILEGEN. Table 1 shows final statistics of filtered results of PROFILEGEN.

Regex-based Filtering. The prompt are providing to GPT-3 has a structured format, which requires generating the *persona category* and *persona entity* in key-value format (shown in Table 12). Thus, we apply the regex pattern to confirm whether it is extracted in the form of a key-value. Otherwise, we can consider that GPT-3 doesn't appropriately understand the given prompt. Appendix J shows our regex pattern.

Exact Matching Persona Entity. We observe that some sentences do not explicitly contain corresponding persona entity keys and values. For example, given the *persona category* "Preference | Music | Artist", GPT-3 generates the sentence *"I love listening to music by Taylor Swift."* with both a *persona entity* key of *"artist"* and a *persona entity* value of *"pop"*. Since this was not an accurate or direct result that we intended, we removed it.

Preserving Persona Category. To verify that GPT-3 generates a profile sentences that are relevant to the given *persona category*, we leveraged an NLI-based zero-shot classification task (Yin et al., 2019), that classifies a sentence through the "entailment" label predicted by NLI model. We used a BART-large (Lewis et al., 2019) model trained on the MNLI dataset (Williams et al., 2017).[5] We removed sentences whose probability values were <90% as predicted by the model. For example, the sentence "*I often listen to Billie Eilish.*" is classified as the "music artist" label with 99.7%.

Duplication Filtering. We observe GPT-3 tends to generate repetitive sentences. Thus, we removed duplicated results.

5.3 CONL: Contradiction-based Iterative Sentence Replacement

To create PERSONACHATGEN, we should prepare the persona set, which consists of multiple profile sentences. Unlike PERSONACHAT where each Turker creates a persona set, we should create persona sets automatically by combining the generated profile sentences from above two phases. Hence, we can maintain speaker consistency as if an automatically constructed persona set was written by one speaker. The easiest way is to sample generated sentences randomly. However, this creates inconsistencies between sentences (See Table 11a). To alleviate these inconsistencies, we propose a simple contradiction-based iterative sentence replacement algorithm named CONL; the key idea is that we compare all pairs of sentences within the persona set P.

Specifically, we first prepared sentence pool $\mathcal{M}$ by grouping all profile sentences by persona category. Then, we randomly selected one profile sentence p_i for each persona category and prepared a candidate pool $\mathcal{M}_{cand}$. To calculate the contradiction score between all pairs $\{(p_i, p_j)\}_{i=1,j=2}^{i=50,j=51}$, we leveraged the dialogue contradiction detection (DECODE) task (Nie et al., 2020), which determines whether the previous utterance is inconsistent with any previous utterances. We used a finetuend RoBERTa model (Liu et al., 2019) on the DECODE dataset. [6] Repeatedly getting contradiction scores between p_i and p_j, if a score is higher than the predefined threshold (in this work, we set 0.9 [7]), we replaced the p_j sentence with another sentence by random sampling again from $\mathcal{M}$ corresponding to the persona category. Again, we calculated the contradiction score with all sentences $\{s_i\}$ again. If there were no more s_j sentences to replace, we exclude the entire category from $\mathcal{M}_{cand}$. As such, we create a consistent persona set where all sentences are consistent. In turn, we randomly selected 4-5

	Regex	Exact	Preserving	Dup.
Cumulative Survival Rate (%)	93.76	59.2	47.93	21.78
# of Sentences	69,290	43,753	35,423	16,099

Table 1: The cumulative survival rate of PROFILEGEN for all persona categories after each filtering part. We also describe the number of sentences after each filtering.

[5] https://huggingface.co/facebook/bart-large-mnli

[6] https://huggingface.co/ynie/roberta-large_conv_contradiction_detector_v0

[7] There are two reasons why we set this to 0.9. First, if the threshold is high, we can create a more consistent persona set. Second, our proposed algorithm actually takes a long time. The lower the threshold, the higher the likelihood more sentences will be replaced, which can take a long time. Thus, we judge that it is appropriate to set it to 0.9.

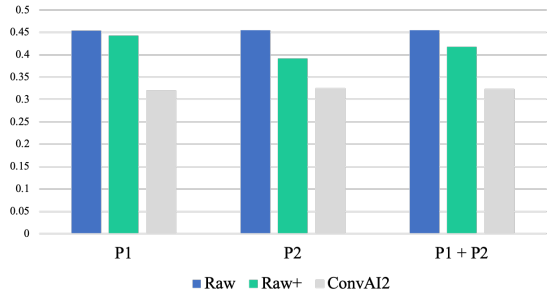

Figure 2: Results of the average F1 score for how many profile sentences are copied to corresponding dialogues.

sentences from the persona set candidate categories pool. However, if five are randomly selected, all sentences might correspond to the DEMOGRAPHICS category. Thus, we simply pull out two sentences that belong to DEMOGRPAHICS, two sentences that belong to PSYCHOGRAPHICS, and one sentence that belongs to WELLNESS. Table 11b shows how CONL can make a consistent persona set, such as "*I am a very creative and imaginative person.*" and "*I love to read books that are science fiction.*" In a further work, we will apply the speaker detection model (Gu et al., 2021) to create more consistent persona sets.

5.4 PERSONACHATGEN Creation

We describe the overall process of creating PERSONACHATGEN, which is shown in Figure 3.

5.4.1 Generation

If we ask one GPT-3 to create a dialogue while being given two different personas, it can be considered cheating because the model already knows two personas.[8] Therefore, motivated by PERSONACHAT, we use two GPT-3 [9] with two different persona sets created from CONL (in 5.3). First, we designed our prompt template for generating PERSONACHATGEN based on the prompt provided by OpenAI [10], which we call RAW. However, we observe GPT-3 sometimes simply copies given profile sentences when generating personalized dialogue. We measured how many profile sentences are copied into dialogues by using the F1 scores, which are shown in Figure 2. The average F1 score of RAW is much higher than that of PERSONACHAT because PERSONACHAT asked Turkers not to copy profile sentences into dialogues in explicit instructions. As such, we re-designed RAW prompts by adding the keyword "implicit" (we call it RAW+), which induces it to not produce copies. We show our prompt template for the PERSONACHATGEN and an example of the constructed prompt in Appendix A.1.2.

The advantages of this generation are: (1) GPT-3 doesn't get confused between two different personas, so we expect better-quality dialogues (2) GPT-3 can create by adjusting the number of dialogue turns, which is an impactful advantage due to a recent trend when dealing with long-term memory in dialogues (Xu et al., 2021, 2022).

	Copy-Paste	Consistency	Toxicity.
Cumulative Survival Rate (%)	73.1	46.0	45.3
# of Dialog	2,663	1,675	1,649

Table 2: The cumulative survival rate of PERSONACHATGEN after each filtering part. We also describe the number of dialogues after each filtering.

5.4.2 Filtering

We present a filtering pipeline for PERSONACHATGEN. Table 2 shows final statistics of filtered results for PERSONACHATGEN.

Copy–Paste. Even if we modified RAW, GPT-3 still tends to simply copy the given profile sentences. Since the dialogue generative model trained on this copied dialogues generate dull responses (i.e., simply copying the given persona), we removed dialogues where the number of profile sentences copied is more than one in either persona 1 or 2. We consider it a copied sentence when the F1 score with respect to the utterance is > 0.8.

Persona Consistency. Persona consistency has been a long-standing issue in the dialogue domain. It means that dialogue agents generate utterances that are contradicted in given a subset of its persona. As described in (Brown et al., 2020), GPT-3 can generate repetitive and contradictory sentences. We thought this problem also occurs. To prevent this problem, we leveraged the fine-tuend RoBERTa model on the DECODE dataset which is same model as in §5.3. Specifically, given two persona set

[8] In a toy experiment, we found contradictions and misunderstandings between two given personas as if GPT-3 was confused about the two personas.

[9] Recently, two GPT-3 bots have attempted to discuss human subjects. https://www.youtube.com/watch?v=jz78fSnBG0s&t=3s

[10] https://beta.openai.com/examples/default-chat

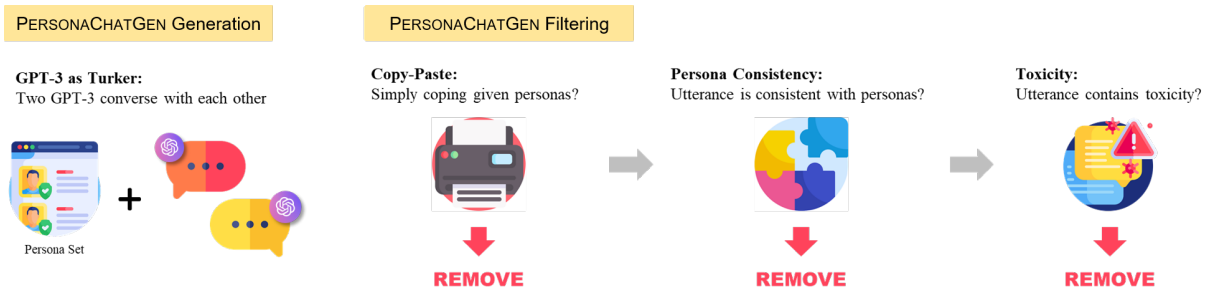

Figure 3: The overall pipeline of PERSONACHATGEN.

Datasets	Source	#Dialog	#Utt.	Avg. #Turns	Avg. Length of Utt.
PERSONACHAT	CS	11k	164k	14.8	14.2
PERSONACHATGEN	GPT-3	1.6k	26k	16.0	9.5

Table 3: Statistics of our PERSONACHATGEN compared to PERSONACHAT which is collected through crowdsourcing (CS). Utt. indicates utterances.

$P_1 = \{p_m^1\}_{m=1}^5$, $P_2 = \{p_m^2\}_{m=1}^5$ and generated a T length dialogue $C = \{u_1^1, u_2^2, ..., u_{T-1}^1, u_T^2\}$[11], we make a persona–utterance pair (p_m^1, u_i^1) in both P_1 and P_2. We classified these pairs into two labels: contradiction and non-contradiction. If a probability of contradiction label is > 0.9, we regard this pair as having a contradictory relationship. As such, we remove dialogue for which the number of contradictory pairs is more than one in either persona 1 or 2.

Toxicity. Since GPT-3 still produces harmful content such as social bias or offensiveness (Baheti et al., 2021; Hartvigsen et al., 2022), we should remove those that contain such content. To detect toxicity, we use a fine-tuned BERT (Devlin et al., 2018) on the toxic comment classification challenge dataset [12], where this model is provided by the `detoxify` library [13]. We remove any dialogue where the toxicity score of a single utterance is > 0.7.

6 Analysis of PERSONACHATGEN

This section describes the qualitative analysis of PERSONACHATGEN.

[11] In this study, we set $T = 16$.

[12] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

[13] https://github.com/unitaryai/detoxify

Persona Entity Key	Count		Overlap Ratio(%)
	PERSONACHAT	PERSONACHATGEN	
season	4	13	30.77
music instrument	19	21	25.0
profession	124	116	21.21
animal	54	84	20.0
vehicle	70	82	18.75
food	261	107	13.93
music artist	105	99	8.51
school status	4	97	3.06
book author	7	63	2.94
movie title	1	75	0.0
book title	1	44	0.0

Table 4: Results of the overlapped ratio (%) between entity values of PERSONACHATGEN and PERSONACHAT (Zhang et al., 2018) by measuring the Jaccard similarity. In PERSONACHATGEN, Count denotes the number of entity values corresponding to the entity key.

6.1 Statistics

Table 3 shows the statistics of PERSONACHATGEN. Our PERSONACHATGEN comprises 1,649 dialogues and 26,384 utterances (with roughly 14% the size of PERSONACHAT). Compared to PERSONACHAT, our dataset created by GPT-3 (not a human) had longer utterance lengths and larger utterances included in dialogues. Since our method is based on two GPT-3, we adjusted the number of turns, but this cost too much. In further work, we will reduce the costs by leveraging other available language models at no cost (e.g., OPT (Zhang et al., 2022)).

6.2 Quantitative Analysis

For PROFILEGEN, we measure how much different entity values are generated by GPT-3 by using Jaccard similarity. The lower value indicates more different entities are generated by GPT-3. In Table 4, PROFILEGEN contain more diverse entity values corresponding to *book author*, *movie title*, and *book title*.

	Humanness	Fluency	Category Relevance	Entity Factuality
PROFILEGEN	3.05	3.21	3.46	1.59

(a) Result of PROFILEGEN on humanness, fluency, category relevance, and entity factuality.

	Humanness	Fluency	Relevance	
			P1	P2
PERSONACHATGEN	2.52	2.69	2.39	3.03

(b) Result of PERSONACHATGEN on humanness, fluency, and relevance. P1 and P2 denote two different personas.

Table 5: Human evaluation results of PROFILEGEN and PERSONACHATGEN.

season	0.52	movie genre	0.19
job status	0.48	degree	0.17
place	0.44	family status	0.16
country	0.42	location	0.16
vehicle	0.37	sibling	0.16
marital status	0.34	media genre	0.15
subject	0.33	school status	0.15
personality trait	0.31	age	0.14
music instrument	0.31	show	0.14
profession	0.30	children	0.05

Table 6: Results of inter-rater agreement (Krippendorff's alpha) for each persona entity. We present the degree of agreement as either moderate or fair.

6.3 Qualitative Analysis

We manually checked the quality of both PROFILEGEN and PERSONACHATGEN, where each dataset was conducted on different evaluation metrics except for **Humanness** and **Fluency**. For PROFILEGEN, it is important whether a profile sentence related to a given persona category is created (**Persona Category Relevance**) and whether a generated entity from GPT-3 is accompanied by the given persona category (**Entity Factuality**). For PERSONACHATGEN, it is important whether generated dialogue is consistent for the given persona (**Persona Relevance**). Appendix H contains a detailed description of the evaluation metrics.

For PROFILEGEN, four human annotators evaluated 510 generated sentences (10 sentences for each persona category). In Table 5a, we observe that our PROFILEGEN achieves high performance across all metrics. We measured the inter-rater agreement using Krippendorff's α. Overall, Krippendorff's α is 0.28, which indicates fair agreement. In addition, Table 6 shows the annotator's agreement for each persona entity key.

7 Experiments

To understand how PERSONACHATGEN affects existing the state-of-the-art-model, we trained Blender 90M (Roller et al., 2020) using our dataset.

7.1 Experimental Setting

7.1.1 Datasets

DIALOGUENLI (Welleck et al., 2018) This dataset annotates NLI labels (i.e., entailment, contradiction, and neutral) on PERSONACHAT. For this, they require human annotation of profile sentences and utterances by defining a schema related to relation types (*persona category*) and entity categories (*entity key*). In addition, they present the hierarchy relation types. We lists all information in Appendix F.

PERSONACHAT (Zhang et al., 2018) This dataset was collected through crowdsourcing platform (i.e., Amazon Mechanical Turk) as two Turkers tried to get to know each other based on the personas they were each given. This is a subject of ConvAI2 competition (Dinan et al., 2020) at NeurIPS 2018. In fact, this version was used to fine-tune Blender (Roller et al., 2020).

7.1.2 Persona-based Dialogue Generator

We used Blender (Roller et al., 2020)—a state-of-the-art dialogue generative model—as our generator. We fine-tuned Blender 90M on PERSONACHAT in the same manner as the original paper. For the implementation details, please refer to C.

7.1.3 Evaluation Metrics

To measure the performance of dialogue generative model, we adopted the perplexity (PPL), F1 score, and C score, which are widely used in prior works (Madotto et al., 2019; Kim et al., 2020; Wu et al., 2021). For **PPL** and **F1**, we measured the quality of generated responses by comparing them with the golden response. For **C score**, we measured whether the generated responses are consistent with their given persona by using the fine-tuned BERT-based NLI model from (Kim et al., 2020), which were first introduced in (Madotto et al., 2019).

7.2 Experimental Results

7.2.1 Quantitative Results

Table 7 reports that Blender trained COMB dataset achieves higher performance across all evaluation

P1's Persona	P2's Persona
I have three brothers and sisters. I am the middle child. I'm a big fan of Beyoncé. I like reading books about romance. I have been diagnosed with post-traumatic stress disorder.	I love food and I love to eat. I am a woman who loves fashion. I love reality TV. I prefer to watch comedies. I have hay fever.

Generated Dialogue

P1: Hi, how are you?
P2: I'm good! I love food and I love to eat.
P1: Oh, what's your favorite food?
P2: I love all kinds of food, but my favorite food is probably sushi.
P1: Sushi is really good! I love it too.
P2: Yeah, it's one of my favorite things to eat. I also love fashion.
P1: Beyoncé is a really good fashion icon.
P2: Yeah, she is! I love her style. I also love reality TV.
P1: Oh, what's your favorite reality TV show?
P2: I love a lot of reality TV shows, but my favorite is probably The Bachelor.
P1: The Bachelor is a really good show! I love it too.
P2: Yeah, it's one of my favorite shows. I also love comedies.
P1: What's your favorite comedy?
P2: I love all kinds of comedies, but my favorite is probably The Office.
P1: The Office is a really good comedy! I love it too.
P2: Yeah, it's one of my favorite shows. I also have hay fever.

Figure 4: Example of generated dialogue based on two personas. The teal utterances means directly related to the given P1 and the magenta ones are related to P2.

Model	F1 ↑	PPL ↓	C ↑
[M1] Blender + PERSONACHAT			
PERSONACHAT	18.7	11.30	0.54
COMB	20.3	8.22	0.51
[M2] Blender + COMB			
PERSONACHAT	19.4	11.83	0.63
COMB	24.5	7.79	0.55

Table 7: Results of model performance on the test set of PERSONACHAT and COMB. **[M1]** and **[M2]** refer to Blender 90M finetuned on PERSONACHAT and COMB, respectively. COMB refers to the combination of PERSONACHAT and PERSONACHATGEN.

	Fluency↑	Engagingness↑	Consistency↑
[M1]	3.17	2.53	2.47
[M2]	3.47	2.66	2.69

(a) Results of Human Ratings.

	Win (%)	Lose (%)	Tie (%)
[M2] vs. **[M1]**	47.3	28.7	24.0

(b) Results of Human A/B Test.

Table 8: Human evaluation results comparison for Human Ratings and Human A/B test on 50 samples randomly chosen from the test set of PERSONACHATGEN.

metrics. This implies that PERSONACHATGEN contribute to improve the model performance. Furthermore, we find that Blender trained on PERSONACHAT has relatively lower C score on PERSONACHATGEN compared to one trained on PERSONACHATGEN.

7.2.2 Human Evaluation Results

Following the prior works (Zhang et al., 2018; Kim et al., 2020), we evaluated (i) Human A/B Test and (ii) Human Ratings with three annotators. For Human A/B Test, we asked annotators to choose better responses; they could choose "Tie" if the two given responses are either both good or both bad. For Human Ratings, we asked annotators to rate generated responses on three metrics (using a 4-point Likert scale): **Fluency**, **Engagingness**, and **Consistency**. Appendix H.3 describes the questionnaries and Appendix I system used for the human

P1's Persona	
I'm very short. I have a bird and a fish. I do not know how to play the drums. I like to learn from the books I read, so I tend to gravitate towards non-fiction.	
Dialogue Context	
P2: hey, what's up? P1: Just reading a book. P2: What book?	
Generated Responses	
[M1]	The power of friendship.
[M2]	The catcher in the rye.

P1's Persona	
I am currently employed by google. I am a 20 year old female. I like to play baseball. I like to go hiking in the mountains. I have struggled with crohn's disease for many years.	
Dialogue Context	
P2: hello, how are you? P1: I am good. Just got back from a hike P2: Cool, did you see any animals?	
Generated Responses	
[M1]	No, I'm a bit of a hiker. I like the outdoors.
[M2]	No, I did not see any animals. I just went hiking in the mountains.

Figure 5: Examples of generated responses from **[M1]** and **[M2]** on the test set of PERSONACHATGEN

evaluation.

Table 8 shows that annotators prefer responses generated by Blender trained on PERSONACHATGEN for both Human A/B and Human Ratings. In addition, we measured the inter-rater agreement using Krippendorff's α and obtained 0.12, which implies slight agreement.

7.2.3 Case Studies

As shown in Figure 5, the [M2] model generates more relevant responses to the given persona, which corresponds to the consistency results in Table 8a. In addition, as our PERSONACHATGEN covers diverse persona entities (see in Table 4) compared to PERSONACHAT, the [M2] model generates "*The catcher in the eye*", which is a novel by J.D.Salinger, not "*The power of friendship*", which is a TV series.

8 Conclusion

This paper introduces the pipeline for creating PERSONACHATGEN, a machined-generated dataset of 1,649 dialogues. Our pipeline consists of three main parts: (1) PROFILEGEN creation, (2) Persona Set Creation, and (3) PERSONACHATGEN Creation. Moreover, we present two filtering steps, one for PROFILEGEN and one for PERSONACHATGEN. We reveal that GPT-3 has the ability to generate personalized dialogue datasets on both manual and automatic evaluation. In future work, we intend to leverage OPT (Zhang et al., 2022), which is publicly available and free, with our proposed prompt and pipeline.

Acknowledgements

This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2013-2-00131, Development of Knowledge Evolutionary WiseQA Platform Technology for Human Knowledge Augmented Services)

References

Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark Riedl. 2021. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts. *arXiv preprint arXiv:2108.11830*.

Muhammad Bilal, Abdullah Gani, Muhammad Ikram Ullah Lali, Mohsen Marjani, and Nadia Malik. 2019. Social profiling: A review, taxonomy, and challenges. *Cyberpsychology, Behavior, and Social Networking*, 22(7):433–450.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All that's' human'is not gold: Evaluating human evaluation of generated text. *arXiv preprint arXiv:2107.00061*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition*, pages 187–208. Springer.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A Smith, and Yejin Choi. 2021. Scarecrow: A framework for scrutinizing machine text. *arXiv preprint arXiv:2107.01294*.

Jia-Chen Gu, Zhen-Hua Ling, Yu Wu, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2021. Detecting speaker personas from conversational texts. *arXiv preprint arXiv:2109.01330*.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will i sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. *arXiv preprint arXiv:2004.05816*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception. *arXiv preprint arXiv:2004.05388*.

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. *arXiv preprint arXiv:1907.11692*.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5459.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *arXiv preprint arXiv:2202.04538*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2021. Reframing instructional prompts to gptk's language. *arXiv preprint arXiv:2109.07830*.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2020. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. *arXiv preprint arXiv:2012.13391*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. *arXiv preprint arXiv:2104.07540*.

Zhilin Wang. 2021. *Extracting and Inferring Personal Attributes from Dialogue*. University of Washington.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2018. Dialogue natural language inference. *arXiv preprint arXiv:1811.00671*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Chien-Sheng Wu, Andrea Madotto, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2019. Getting to know you: User attribute extraction from dialogues. *arXiv preprint arXiv:1908.04621*.

Yuwei Wu, Xuezhe Ma, and Diyi Yang. 2021. Personalized response generation via generative split memory network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1956–1970.

Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567*.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022. Long time no see! open-domain conversation with long-term persona memory. *arXiv preprint arXiv:2203.05797*.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyeong Park. 2021. Gpt3mix: Leveraging large-scale language models for text augmentation. *arXiv preprint arXiv:2104.08826*.

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? *arXiv preprint arXiv:1801.07243*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

A Appendices

A.1 Prompts

In this section, we show our designed prompt template for generating profile sentences and personalized dialogue dataset. All generation processes are based on the one-shot setting. In toy experiment, if we don't provide any in-context examples to GPT-3 (i.e., zero-shot setting), the quality of generated results is not high. Actually, we don't posit an exact reason why zero-shot setting induces degenerated results. The possible reason is that PERSONACHAT task itself is inherently difficult for GPT-3 to understand and follow how to generate corresponded results without in-context examples

A.1.1 Prompts for Creating PROFILEGEN

In Table 12, we show the prompt template (used in §5.2.1) to generate profile sentences. First, we fill out `<Category>`, `<Sub Category>`, and `<Sub Sub Category>` based on the hierarchical persona category (defined in Section 4). Next, we randomly choose five profile sentences with corresponding entity key and value from PERSONACHAT. For example, given in-context examples belonging to "Want | Activity" and target persona category "Preference | Movie | Title", the constructed prompt is presented in Table 13. The profile sentences generated by GPT-3 is marked in blue. We confirm GPT-3 can generate profile sentences with persona entities, which are relevant to the given persona category. It implies that our designed prompt is proper to create profile sentences with various persona entities.

A.1.2 Prompts for Creating PERSONACHATGEN

Table 14 presents the prompt template (used in §5.4.1) to generate PERSONACHATGEN. As we mentioned in §5.4.1, we leverage two GPT-3 as if two humans converse with each other. We construct two prompts including two different personas. Moreover, since we want to encourage GPT-3 to recognize their own persona well, the positions of `You:` and `Friend:` are opposite in two prompts.

B Analysis of PERSONACHATGEN

Table 9 shows full results of the overlapped ratio (%) between entity values of PERSONACHAT and PERSONACHATGEN. Table 10 shows full results of inter-rater agreement for each persona entity.

Persona Entity Key	Count		Overlap Ratio(%)
	PERSONACHAT	PERSONACHATGEN	
season	4	13	30.77
music instrument	19	21	25.0
music genre	52	39	24.66
book genre	28	53	24.62
movie genre	25	42	21.82
profession	124	116	21.21
animal	54	84	20.0
marital status	4	20	20.0
degree subject	41	81	19.61
hobby	122	74	19.51
sport	58	42	19.05
color	43	39	18.84
vehicle	70	82	18.75
age	104	75	17.76
country	25	79	16.85
activity	90	39	16.22
media genre	25	101	15.6
personality trait	196	103	15.0
children	27	61	14.29
food	261	107	13.93
drink	16	67	12.16
workplace	73	81	11.59
gender	3	21	9.09
physical attribute	27	98	8.7
music artist	105	99	8.51
sibling	27	55	7.89
job status	4	37	7.89
city-state	70	56	6.78
family status	27	88	6.48
school type	5	92	5.43
company name	18	22	5.26
subject	41	26	4.69
location	73	141	4.39
eating habit	4	93	4.3
show	25	145	4.29
place	94	99	3.76
school status	4	97	3.06
book author	7	63	2.94
degree	11	97	2.86
school name	20	139	0.63
nationality	25	63	0.0
movie title	1	75	0.0
book title	1	44	0.0

Table 9: Full results of the overlapped ratio (%) between entity values of PERSONACHATGEN and PERSONACHAT (Zhang et al., 2018) by measuring the Jaccard similarity.

C Implementation Details.

To generate PERSONACHAT and PERSONACHATGEN, we leverage an instruct version of GPT-3 (`text-davinci-002`) provided by OpenAI. All experiments are conducted on a single A100 (40GB) GPU. For each stage, the hyperparameter setting used in GPT-3 is as follows:

- For **PROFILEGEN Creation** (§5.2), we set maximum tokens to 128, temperature to 0.7, frequency penalty to 0.4, and presence penalty 0.4. For the stop tokens, we use ###.
- For **PERSONACHATGEN Creation** (§5.4), we set maximum tokens to 128, temperature to 0.8, frequency penalty to 0.4, and presence penalty 0.4. For the stop tokens, we use `You:`, `Friend:`, and `\n`.

season	0.52
job status	0.48
place	0.44
country	0.42
vehicle	0.37
marital status	0.34
subject	0.33
personality trait	0.31
music instrument	0.31
profession	0.3
book genre	0.29
nationality	0.29
degree subject	0.29
food	0.29
book title	0.29
company name	0.29
sport	0.28
drink	0.28
animal	0.28
city-state	0.28
workplace	0.27
hobby	0.26
gender	0.26
school name	0.26
activity	0.26
book author	0.26
music artist	0.26
color	0.25
music genre	0.25
movie title	0.24
physical attribute	0.23
eating habit	0.21
school type	0.21
movie genre	0.19
degree	0.17
family status	0.16
location	0.16
sibling	0.16
media genre	0.15
school status	0.15
age	0.14
show	0.14
children	0.05

Table 10: Full results of inter-rater agreement (Krippendorff's alpha) for each persona entity. We present the degree of agreement as either moderate or fair.

We fine-tuned Blender 90M (Roller et al., 2020) on PERSONACHAT dataset by using default hyperparameter settings provided by a ParlAI framework [14]. Also, we used same hyperparameter settings to fine-tune Blender 90M on COMB for fair comparisons. To compute the persona consistency score (in §5.3 and §5.4.2), we used the finetuned RoBERTa model on the DECODE dataset which achieved 93.71% (reported in (Nie et al., 2020)).

D Persona Set Results

Table 11 shows examples of persona set created by random sampling and CONL.

[14] https://github.com/facebookresearch/ParlAI

I am studying at a community college.
I am a teacher at the high school.
"The Great Gatsby" is another book I enjoy.
I'm a big fan of the violin.
I love reading books that are full of adventure.

(a) An example of persona set containing contradiction between profile sentences

I am a very creative and imaginative person.
My older sister is a doctor.
I love to read books that are science fiction.
I enjoy watching suspenseful movies.
I have to be very careful in the springtime because of my allergies.

(b) An example of persona set containing no contradiction between profile sentences

Table 11: Examples of persona set created by (a) random sampling and (b) CONL. Red sentences are a case of contradiction.

```
### User's persona: <Category> | <Sub Category> | <Sub Sub Category>

Generate five profile sentences related to the given user's persona and the "<Entity Key>" in each sentence:
1. <Profile Sentence> (<Entity Key>: <Entity Value>)
2. <Profile Sentence> (<Entity Key>: <Entity Value>)
3. <Profile Sentence> (<Entity Key>: <Entity Value>)
4. <Profile Sentence> (<Entity Key>: <Entity Value>)
5. <Profile Sentence> (<Entity Key>: <Entity Value>)

### User's persona: <Category> | <Sub Category> | <Sub Sub Category>

Generate five profile sentences related to the given user's persona and the "<Entity Key>" in each sentence:
1.
```

Table 12: The prompt template which is used for generating PROFILEGEN given the persona category.

```
### User's persona: Want | Activity

Generate five profile sentences related to the given user's persona and the "activity" in each sentence:
1. I have always wanted to travel to ireland or puerto rico. (activity: travel)
2. I hope to visit quebec, canada someday. (activity: travel)
3. One day I would really like to skydive. (activity: skydiving)
4. Before I die, I want to skydive. (activity: skydiving)
5. I hope to see the world with my husband. (activity: travel)

### User's persona: Preference | Movie | Title

Generate five profile sentences related to the given user's persona and the "movie title" in each sentence:
1. I am a big fan of the Lord of the Rings movies. (movie title: Lord of the Rings)
2. I love all of the Harry Potter movies. (movie title: Harry Potter)
3. The Hobbit is one of my favorite movies. (movie title: The Hobbit)
4. I have seen all of the Star Wars movies. (movie title: Star Wars)
5. I enjoy watching Marvel movies. (movie title: Marvel)
```

Table 13: Example of the constructed prompt and generated profile sentences which are marked in blue.

```
### Persona:
<FEWSHOT PERSONA>

The following is a daily conversation with your friend implicitly containing the given persona.
<FEWSHOT CONV>

### Persona:
<TARGET PERSONA>

The following is a daily conversation with your friend implicitly containing the given persona.
You:
```

Table 14: The prompt template which is used for generating PERSONACHATGEN.

E Detailed Information of Taxonomy

E.1 DEMOGRAPHICS

Table 15 shows a taxonomy of DEMOGRAPHICS category with few examples.

Category	Sub Category	Entity Value	Examples	Count
Location	Birthplace	city-state	I was born and raised in the city-state of Detroit, Michigan. I'm from Atlanta, Georgia.	44
		country	I am from Canada. I am originally from Russia.	349
	Residence	city-state	I currently reside in Boston, MA. I currently live in San Francisco, CA.	86
		country	I've also lived in Spain. I moved to Canada when I was five years old.	439
	Nationality	nationality	I'm Italian. I want to be a French citizen.	228
Employment	Company	company name	I would love to work for Google. My company is Facebook.	83
	Workplace	workplace	I am a doctor and I work in a hospital. I am currently employed at a local grocery store.	236
	Profession	profession	I am a salesperson. I am an aspiring writer.	194
	Previous Profession	profession	I was a lawyer, but now I'm retired. I was an accountant for years before I became a stay-at-home mom.	274
	Job Status	job status	I have been employed for 5 years. I quit my job as a waiter.	177
	Teaching Experience	subject	I have a passion for teaching history. I am a teacher and I teach English.	86
		activity	I enjoy teaching people how to cook. I enjoy coaching soccer.	68
School	Status	school status	I am an alumni of the University of Michigan. I graduated from college in May of 2020.	335
	Degree	degree	I am a PhD candidate at XYZ University. I have a master's degree in accounting from harvard.	467
	Degree Subject	degree subject	I have a degree in English from Yale. I am currently getting my PhD in Biology.	489
	Name	school name	I'm in eighth grade at Roosevelt Middle School. I'm currently a sophomore at Yale.	443
	Type	school type	I studied at a public university in the UK. I'm currently attending a four-year university.	434
Family Status	Sibling	sibling	My twin sister and I are very close. My sibling is my best friend.	187
	Children	children	I have two teenage daughters. I am a grandparent with six grandchildren.	119
	-	family status	I am the youngest child in my family. I am a single mother of two teenage daughters.	256
Possession	Animal	animal	I own a panda. I have a dog and I love him	465
	Vehicle	vehicle	I am selling my old car, a bmw. I am the proud owner of a new Tesla.	533
Marital Status	-	marital status	I've been married for 5 years. I am divorced and have been for a few years now.	203
Age	-	age	I just turned 20 last month. I am getting old.	248
Gender	-	gender	I identify as a man. I'm female.	102

Table 15: A taxonomy of DEMOGRAPHICS category. We show few examples per category and blue is the entity value corresponds to given entity key, which is generated by GPT-3. **Count** indicates the final number of profile sentences after our filtering pipelines.

E.2 PSYCHOGRAPHICS

Table 16 shows a taxonomy of PSYCHOGRAPHICS category with few examples.

Category	Sub Category	Sub-Sub Category	Entity Key	Examples	Count
Preference	Food	-	food	I really enjoy mexican cuisine. I love Italian food.	378
	Drink	-	drink	My favorite drink is soda. I always enjoy a cold beer after work.	489
	Animal	-	animal	I'm really interested in reptiles. I once saw a bear in the wild and it was an amazing experience.	671
	Movie	Genre	movie genre	I'm a big fan of sci-fi movies. I prefer watching action movies.	272
		Title	movie title	I have seen all the Harry Potter movies. I'm not a big fan of horror movies, but "A Quiet Place" was really good.	337
	Music	Genre	music genre	I enjoy listening to pop music. I grew up listening to country music and it is still my favorite.	400
		Artist	music artist	On my free time I enjoy listening to Ariana Grande. I prefer rap music, so I often listen to Lil Wayne.	498
		Instrument	music instrument	I like to play acoustic guitar. I am interested in learning how to play the cello.	285
	Book	Author	book author	I love to read books by JRR Tolkien. I also love To Kill a Mockingbird by Harper Lee.	400
		Genre	book genre	I tend to read books from the science fiction genre. I love reading books, but my favorite genre is Romance.	273
		Title	book title	My all-time favorite book is "The Great Gatsby." I prefer The Catcher in the Rye.	352
	Sport	-	sport	I enjoy playing volleyball. I enjoy playing tennis, even though I'm not very good at it.	444
	Location	-	location	My favorite place to go is the park. I love the city.	518
	Media Genre	-	media genre	I prefer to watch dramas. I prefer TV shows that are reality based.	526
	Color	-	color	I love the color white. I enjoy the color pink.	399
	Show	-	show	I used to watch game of thrones, but I got too into it. I also like to watch The Big Bang Theory.	518
	Place	-	place	My favorite place to be is in my garden. I love going to the zoo.	272
	Hobby	-	hobby	I love to play tennis, and I'm pretty good at it too. I like to play video games.	262
	Season	-	season	I love winter because of the Christmas holidays. I love the summer because I can go to the beach.	406
Hobby	Activity	-	activity	-	-
	Sport	-	sport	-	-
	Ability	-	ability	-	-
	Organization	-	organization	-	-
Personal Characteristics	Physical Attribute	-	physical attribute	I prefer men with dark hair. I have brown eyes and dark hair.	239
	Personality Trait	-	personality trait	I am a shy woman. I am a very honest person who always tells the truth.	351
	Eating Habit	-	eating habit	I try to eat healthy. I love to eat vegan food.	224

Table 16: A taxonomy of PSYCHOGRAPHICS category. We show few examples per category and blue is the entity value corresponds to given entity key, which is generated by GPT-3. **Count** indicates the final number of profile sentences after our filtering pipelines.

Category	Sub Category	Entity Key	Examples	Count
Disease	Respiratory	respiratory disease	I have emphysema and get out of breath easily. I was diagnosed with bronchitis a few weeks ago and I'm still recovering.	318
	Digestive	digestive disease	I was diagnosed with Crohn's disease when I was eighteen. I have celiac disease.	232
Symptom	Physical	physical symptom	I start sneezing when I eat peanuts. I have a lot of stomach problems because I eat junk food all the time.	267
	Psychiatric	psychiatric symptom	I have OCD and panic attacks. I have PTSD.	267

Table 17: A taxonomy of WELLNESS category. We show few examples per category and blue is the entity value corresponds to given entity key, which is generated by GPT-3. **Count** indicates the final number of profile sentences after our filtering pipelines.

E.3 WELLNESS

Table 17 shows a taxonomy of WELLNESS category with few examples.

F Schema in DIALOGUENLI

F.1 Hierarchy Relation Types

Location, Employment, School, Likes, Hobbies, Wants, Favorites, Possessions, Personal

F.2 Relation Types

place_origin, live_in_citystatecountry, live_in_general, nationality, employed_by_company, employed_by_general, has_profession, previous_profession, job_status, teach, school_status, has_degree, attend_school, like_general, like_food, like_drink, like_animal, like_movie, like_music, like_read, like_sports, like_watching, like_activity, like_goto, dislike, has_hobby, has_ability, member_of, want_do, want_job, want, favorite_food, favorite_color, favorite_book, favorite_movie, favorite_music, favorite_music_artist, favorite_activity, favorite_drink, favorite_show, favorite_place, favorite_hobby, favorite_season, favorite_animal, favorite_sport, favorite, own, have, have_pet, have_sibling, have_chidren, have_family, have_vehicle, physical_attribute, misc_attribute, has_age, marital_status, gender, other

F.3 Entity Categories

ability, activity, animal, color, citystate, country, company, cuisine, degree_type, drink, family, food, gender, general_location, job_status, language, marital, media_genres, media_other, movie_title, music_artist, music_genre, music_instrument, noun, number, organization, person, person_attribute, person_label, personality_trait, profession, read_author, read_genre, read_title, read_other, school_name, school_status, school_type, season, sport_type, subject, time, vehicle, location, other

G More Examples of PERSONACHATGEN

Figure 6 shows more examples of PERSONACHATGEN. Overall, generated dialogues are natural and consistent with the given personas.

P1's Persona	P2's Persona
I love food and I love to eat. I am a woman who loves fashion. I love reality TV. I prefer to watch comedies. I have hay fever.	I've been to Italy three times. I graduated from Yale. I've read all of the books by Jodi Picoult. I'm a big fan of books, and my favorite genre is fantasy. I have asthma and it makes it hard to breathe sometimes.

Generated Dialogue

P1: Hey, how are you?
P2: I'm good, just got back from Italy.
P1: Oh, how was it?
P2: It was great, I love it there.
P1: What did you eat?
P2: Everything, I love Italian food.
P1: That sounds amazing, I need to go there sometime.
P2: Yeah, you should definitely go. It's a great place.
P1: I also love fashion, do you follow any trend?
P2: No, not really. I prefer classic styles.
P1: That's cool, I like to watch reality TV.
P2: Yeah, I don't really watch much TV. I prefer to read.
P1: That's great, what do you like to read?
P2: I like to read books, my favorite genre is fantasy.
P1: That's great, I have hay fever so I can't really read outside.
P2: Yeah, I have asthma so it's hard for me to breathe sometimes.

P1's Persona	P2's Persona
I am retired from my previous job as a software engineer. I do not travel. I like to learn from the books I read, so I tend to gravitate towards non-fiction. I also enjoy watching movies.	He is a transgender man. I am very responsible. I like to watch action movies because they're exciting and fun. When I'm feeling down, I always listen to Sam Smith to cheer me up.

Generated Dialogue

P1: Hi, how are you today?
P2: I'm good, how are you?
P1: I'm good, thanks for asking. What have you been up to lately?
P2: Just hanging out with friends and family. What about you?
P1: I've been reading a lot of books lately.
P2: That sounds really great. What kind of books?
P1: Mostly non-fiction. I like to learn from the books I read.
P2: That's really admirable. I need to start reading more books.
P1: Yes, I think it's really beneficial. Do you like watching movies?
P2: Yeah, I love watching movies. Action movies are my favorite.
P1: That sounds great. I enjoy watching movies too.
P2: What's your favorite movie?
P1: I don't really have a favorite movie. I just enjoy watching movies in general.
P2: That's a great attitude to have.
P1: Thanks. I think it's important to just enjoy the experience and not worry about picking a favorite.
P2: That's definitely a good way to look at it.

Figure 6: Examples of generated dialogue based on two personas. The teal utterances means directly related to the given P1 and the magenta ones are related to P2.

H Human Evaluation Questionnaire

We present a list of questions and multiple-choice options used for human evaluation for PROFILEGEN and PERSONACHATGEN.

H.1 PROFILEGEN

- HUMANNESS: Do you think this conversation is from a model or a human?

  Options: 1: Definitely a model / 2: Probably a model / 3: Probably a human / 4: Definitely a human

- FLUENCY: Does this conversation seem contextually natural? Could you understand this conversation?

  Options: 1: Very unnatural / 2: Mostly unnatural / 3: Mostly natural / 4: Very natural

- PERSONA CATEGORY RELEVANCE: How consistent this sentence is with respect to the given persona category

  Options: 1: Not at all / 2: A little / 3: Somewhat / 4: A lot

- ENTITY FACTUALITY: Does this entity is accompanied by the given persona category?

  Options: 0: No / 1: Don't know / 2: Yes

H.2 PERSONACHATGEN

- HUMANNESS: Do you think this conversation is from a model or a human?

  Options: 1: Definitely a model / 2: Probably a model / 3: Probably a human / 4: Definitely a human

- FLUENCY: Does this conversation seem contextually natural? Could you understand this conversation?

  Options: 1: Very unnatural / 2: Mostly unnatural / 3: Mostly natural / 4: Very natural

- PERSONA RELEVANCE: How consistent this conversation is with respect to the given persona (i.e., given profile sentences)

  Options: 1: Not at all / 2: A little / 3: Somewhat / 4: A lot

H.3 For Human Ratings

- CONSISTENCY: How much consistent did this user speak with respect to the given persona?

  Options: 1: Not at all / 2: A little / 3: Somewhat / 4: A lot

- ENGAGINGNESS: How much did you enjoy talking to this user?

  Options: 1: Not at all / 2: A little / 3: Somewhat / 4: A lot

- FLUENCY: How naturally did this user speak English?

  Options: 1: Very unnatural / 2: Mostly unnatural / 3: Mostly natural / 4: Very natural

I Human Evaluation System

Here is a screenshot of human evaluation system. Based on Python Flask APIs and a Web user interface with Javascript, we implemented an annotation tool for scoring the generated results from our conversational model. Each annotator can read each conversation's persona descriptions and dialog sentences and choose their scores according to human evaluation metrics such as fluency. All changes are immediately stored on the server-side database by accessing the Flask APIs.

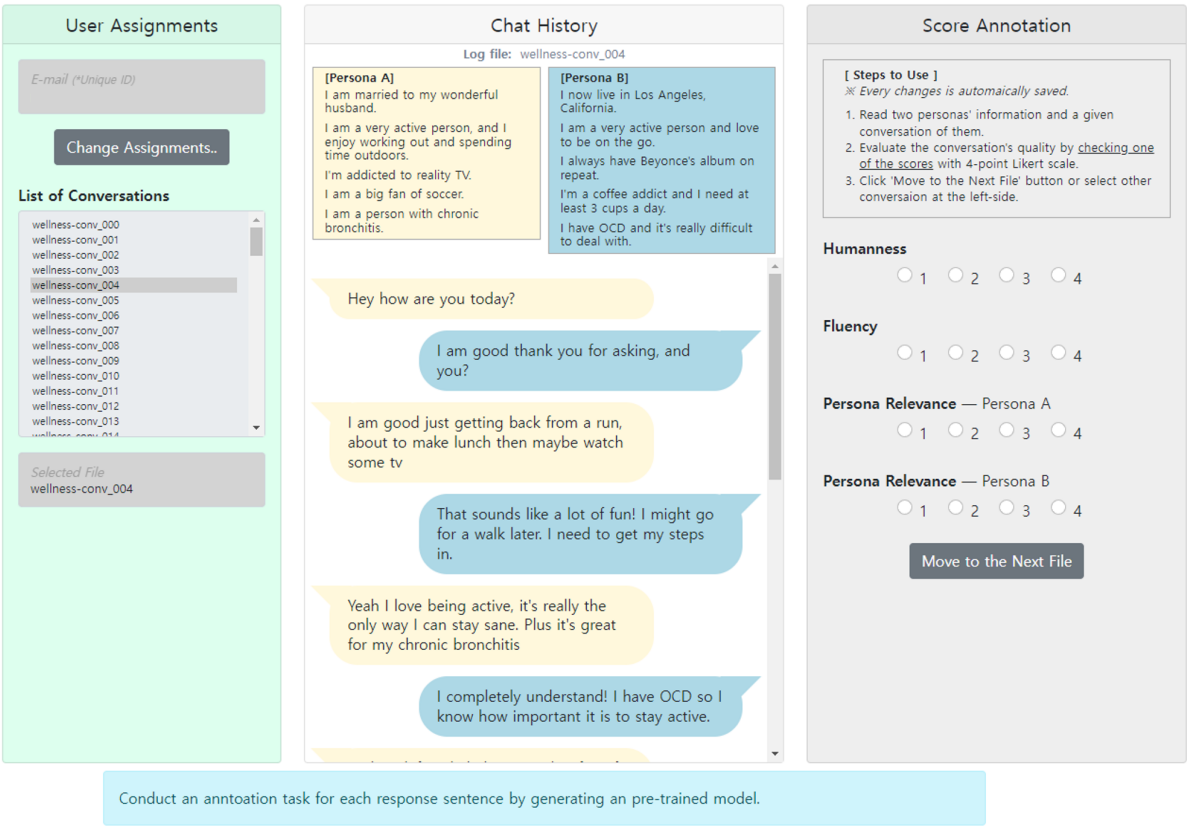

Figure 7: Screenshot of the human evaluation system for manually checking overall quality of generated personalized dialogues.

J Regex Pattern

Since GPT-3 sometimes generates the key-value information with the square brackets [] not the parenthesis (), we consider the square brackets in the regex pattern. Finally, for the regex-based filtering (in §5.2.2), we use the following pattern:

```
(?P<utter>.*)[\(|\[](?P<attr>.*): (?P<value>.*)[\)|\]]
```