# Faithful Persona-based Conversational Dataset Generation with Large Language Models

**Pegah Jandaghi** *
University of Southern California
jandaghi@usc.edu

**XiangHai Sheng**
Google
xhs@google.com

**Xinyi Bai**
Google
shinii@google.com

**Jay Pujara**
Information Sciences Institute
jpujara@isi.edu

**Hakim Sidahmed**
Google Research
hsidahmed@google.com

## Abstract

High-quality conversational datasets are essential for developing AI models that can communicate with users. One way to foster deeper interactions between a chatbot and its user is through *personas*, aspects of the user's character that provide insights into their personality, motivations, and behaviors. Training Natural Language Processing (NLP) models on a diverse and comprehensive persona-based dataset can lead to conversational models that create a deeper connection with the user, and maintain their engagement. In this paper, we leverage the power of Large Language Models (LLMs) to create a large, high-quality conversational dataset from a seed dataset. We propose a Generator-Critic architecture framework to expand the initial dataset, while improving the quality of its conversations. The Generator is an LLM prompted to output conversations. The Critic consists of a mixture of expert LLMs that control the quality of the generated conversations. These experts select the best generated conversations, which we then use to improve the Generator. We release Synthetic-Persona-Chat[1], consisting of 20k conversations seeded from Persona-Chat (Zhang et al., 2018). We evaluate the quality of Synthetic-Persona-Chat and our generation framework on different dimensions through extensive experiments, and observe that the losing rate of Synthetic-Persona-Chat against Persona-Chat during an AI detection test decreases from 17.2% to 8.8% over three iterations.

## 1 Introduction

Every person is a story. Systems that interact with people must understand their underlying stories to effectively engage with them. Unfortunately, many existing datasets used for training conversational agents do not sufficiently model their users. *Personas* - abstract user representations that express the "story" of a person based on their background and preferences - have been widely used for human-centered design in a variety of domains, including marketing, system design, and healthcare (Pruitt and Grudin, 2003b). Prior persona-based conversational datasets, like Persona-Chat (PC) (Zhang et al., 2018), suffer from several limitations, such as small size, static dialogues that cannot easily be updated with new topics, irrelevant utterances, and contradictory *persona attributes* (Wu et al., 2019). In this paper, we propose a novel framework for generating large, dynamic, persona-based conversational datasets that capture the breadth and depth of human experience.

Personas (Pruitt and Grudin, 2003a; Cooper and Saffo, 1999) have been widely used in a variety of domains and applications, including creating narratives for patients and sharing educational messages in healthcare (Massey et al., 2021), targeting users in marketing (van Pinxteren et al., 2020; Fuglerud et al., 2020), and communicating with workers in management (Claus, 2019). Conversational agents use personas to generate more interesting and engaging conversations with their users (Zhou et al., 2020; Shum et al., 2019).

Creating persona-based datasets is difficult: the process is labor-intensive, the outputs must be updated to reflect current events and new concepts, and there are often quality concerns. Existing persona-based datasets have resulted from labor-intensive data collection processes (Zhang et al., 2018; Zhong et al., 2020) involving humans to create or validate personas, create fictional persona-based conversations, and ensure the conversations are coherent. Moreover, even after these datasets are created, it is difficult to update them with the latest topics (Lee et al., 2022), such as current events, new concepts, products, or social trends (Lazaridou et al., 2021). Finally, existing persona-based

* Work done during an internship at Google Inc., Mountain View, USA

[1] Dataset will be publicly available on Github

datasets do not guarantee *faithfulness*, a criterion we introduce to describe the alignment between participants' utterances and their personas.

In this paper, we introduce a new framework for generating large, customized persona-based conversational datasets that uses unsupervised LLMs to reduce human labor, introduces methods to generate, expand, and update personas automatically, and enforces a set of quality criteria including faithfulness to ensure dialogues are human-like. Our persona-based conversational dataset generation framework consists of a three-level pipeline:

1. User Generation
2. User Pairing
3. Conversation Generation

The user generation step takes a set of seed personas, and augments it to create plausible user profiles. The user pairing step matches users to participate in conversations. The conversation generation produces plausible conversations between the selected user pairs. The conversation generation component uses a method similar to self-feedback (Madaan et al., 2023) to iteratively improve the quality of generated samples.

We used the proposed framework to create Synthetic-Persona-Chat (SPC), a conversational dataset with $5k$ user personas, and $20k$ faithful dialogues. The framework we defined to create this dataset can be reused to define specialized personas, such as user music profiles, etc. to create application-specific datasets.

Our contributions are:

- We propose an unsupervised approach to generate, and extend specialized personas using LLMs.
- We introduce and evaluate a framework based on LLMs to evolve a dataset while imposing different objectives on it.
- We release Synthetic-Persona-Chat, a high-quality, faithful, persona-based conversational dataset useful for several conversational tasks, such as training persona inference models.

## 2 Definitions

We define the faithful persona-based dialogue generation task. We begin by defining the persona-based dialogue generation task. We then formally define the faithfulness criteria as a desired quality for the generated dialogues. Throughout this section, we use $\pi$ to refer to persona attributes (individual sentences which, together, form the user persona), $U$ to refer to user profiles, and $D$ to refer to conversations (dialogues).

**Persona Attributes** We define a user persona attribute as a sentence describing this user. "I like ice cream", "I have two brothers" and "My native language is Tamazight" are all examples of persona attributes. Let $\Omega$ be the universal set of persona attributes. $\Omega$ contains all natural language descriptions of all tangible features of any person, which is unbounded.

**Persona Categories** To help organize the vast space of personas, we adopt the approach of Lee et al. (2022) who introduced persona categories. Persona categories are groups of persona attributes that describe the same semantic feature of the user. In our work, we associate each persona category with a corresponding query that can be answered with all persona attributes in that category. For example, job and family situation are persona categories, and corresponding queries might be "What is your occupation?", and "Do you have a family?".

**Persona Attribute Structure** Persona attributes can overlap. For instance, the attribute "I introduced my kids to scuba diving at a young age" overlaps with the attribute "My eldest son goes to elementary school", since both include the "parenthood" feature of the user. Moreover, some persona attributes form a hierarchy, and some persona attributes are specific cases of other attributes.

**User Profile** We define a user profile as a set of persona attributes that can be used to describe a user. For a realistic user, the persona attributes describing a user profile should not contradict each other, and be consistent. An arbitrary persona attribute set $U \subset \Omega$ is a consistent set of persona attribute if, and only if:

$$\forall \pi_1 \in U, \nexists \Pi_2 \subset U : (\Pi_2 \neq \varnothing) \wedge (\Pi_2 \rightarrow \neg \pi_1)$$

**Persona-based Conversation** A persona-based conversation $D$ contains utterances such that at least one persona attribute from each user profile can be inferred from it. For example, the persona attribute "I am a parent" can be inferred from the utterance "I just dropped off my son at school". A persona-based conversation model is a generative model that takes a pair of user profiles ($U_1$, $U_2$) as input, and returns a persona-based dialogue $D$ between these two users.

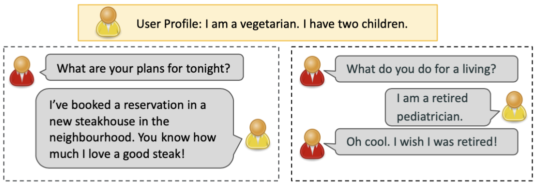



Figure 1: Unfaithful Conversation (Left): Loving steak is negatively correlated with the persona attribute "I am a vegetarian". Faithful Conversation (Right): It introduces no information that contradicts or weakens the user's profile.

**Faithfulness** One crucial quality for a persona-based conversation is that it should align with the user profile. Inspired by (Daheim et al., 2023) which introduces dialogue system faithfulness to the knowledge contained in relevant documents, we specify the criterion of *faithfulness* to characterize the alignment between the utterances of a user in a persona-based conversation and their profile. The faithfulness criterion enforces the constraint that the utterances of a user should not decrease the likelihood of their persona. This criterion assumes the existence of both a prior probability of persona attributes, and an inference model for determining the probability of persona attributes conditioned on utterances. Let $M$ be such an inference model, ($U_1$, $U_2$) a pair of user profiles, and $D$ a persona-based conversation between them. To be a faithful conversation based on $M$, $D$ should not contain any contradicting evidence to the persona attributes of the speakers: passing the conversation $D$ as input to the inference model $M$ should not reduce the inference probability of persona attributes in either of the user profiles $U_1$ or $U_2$. In other words, the probability of any persona attribute in the user profiles based on conversation $D$ should not be less than the probability of that persona attribute without any assumptions. Formally, we call a conversation $D$ faithful with respect to the user profiles $U_1$ and $U_2$, and inference model $M$ if the following condition holds: $\forall \pi \in U_1 \cup U_2 : P_M(\pi|D) \geqslant P_M(\pi)$. Where $P_M(\pi|D)$ indicates the probability that $M$ infers the persona $\pi$ given conversation $D$. We show examples of faithful, and unfaithful conversations in Figure 1.

# 3 Method

In this section, we introduce our method to generate persona-based conversations. We create such conversations with minimum human input, starting from an initial dataset. Our process consists of three steps, as shown in Figure 2: user generation, user pairing, and conversation generation.

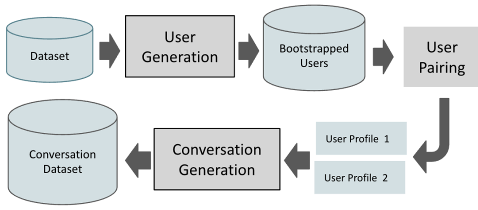



Figure 2: Dataset Augmentation Pipeline

The first component augments a set of seed persona attributes $\Pi_0$ into an expanded set of persona attributes $\Pi_e$, from which it creates user profiles. The second component pairs user profiles as interlocutors of a conversation. The third and final component uses an iterative process to generate high-quality conversations among user profile pairs. We detail each of these components below.

## 3.1 User Generation

The User Generation component is split into two sub-components:

1. Persona Expansion
2. User Profile Construction

We bootstrap seed persona attributes by using various prompts (Brown et al., 2020a) to generate new persona attributes in the Persona Expansion step (Refer to Appendix A.1 for more details on the prompts used). We then create new user profiles by iteratively selecting random user persona attributes from the expanded persona attributes. We employ a Natural Language Inference (NLI) model to ensure the consistency of the constructed user profiles.

### 3.1.1 Persona Expansion

We propose an unsupervised method to augment a set of seed persona attributes $\Pi_0$ into a super-set $\Pi_e$. Unlike previous approaches (Lee et al., 2022), our method is independent of human knowledge or intervention, making it capable of creating specialized personas in new domains. We proceed in two steps: query induction, and persona bootstrapping. In the query induction phase, we identify persona categories in $\Pi_0$, along with associated queries. We then expand these queries into a set $Q$ that also covers unobserved persona categories. The persona bootstrapping step leverages the category-based query set $Q$, and the initial persona attribute seed set $\Pi_0$ to generate new persona attributes. Both of these steps are based on the bootstrapping technique (Yarowsky, 1995), and involve prompting an LLM. We provide a detailed description of these two steps in the following.

**Query Induction** As described in Section 2, each persona attribute belongs to at least one persona category, and each category is associated with a corresponding query that can be answered with persona attributes in that category. The query induction process initially identifies the queries associated with persona categories in $\Pi_0$. It then bootstraps queries by feeding them to a prompted LLM to create more queries that are associated with unobserved categories, ultimately creating a query set $Q$. Including queries associated with unobserved persona categories facilitates the creation of a more diverse set of personas, and increases the scale of augmentation.

The query induction relies on the following assumption:

**Assumption** *Let $\mathcal{M}$ be an LLM, and let $\Gamma$ be the set of all queries associated with all persona categories. If two persona attributes $\pi_1$ and $\pi_2$ belong to the same persona category, then there exists a query $q^{\mathcal{M}} \in \Gamma$ such that $\pi_1$ and $\pi_2$ are $\mathcal{M}$'s output to $q^{\mathcal{M}}$.*

The persona attributes "I am a doctor" and "I am a truck driver", for instance, both belong to the "job" category, leading to the query "What is your job?". We use an agglomerative clustering method to identify the persona categories in $\Pi_0$. Let $C$ be an arbitrary persona cluster in $\Pi_0$. To generate a query for $C$, we select a random subset of persona attributes in $C$, and create a prompt using these samples. We employ this strategy to generate queries for all the clusters identified in $\Pi_0$, and create a set of queries, which we refer to as $Q_0$. Details on the clustering, query induction, together with examples of clusters, persona attributes, and induced queries are available in Appendix A.1. We come up with queries for new, unobserved persona categories by bootstrapping the queries in $Q_0$: starting from $Q = Q_0$, we iteratively sample a set of queries from $Q$, and create a prompt by concatenating them. We then prompt the LLM to generate a new query, and add it to the query set $Q$, as shown in Figure 3. We generated a total of $|Q| = 188$ queries. This set of category-specific queries $Q$ is later used to guide the LLM to generate new persona attributes from the specified category. Thus, higher values of $|Q|$ result in greater diversity within the expanded persona attribute set.

**Persona Bootstrapping** We use the persona attribute seed set $\Pi_0$ and category-specific queries

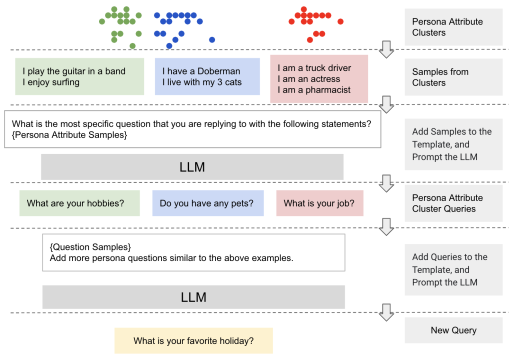



Figure 3: Query Induction Steps

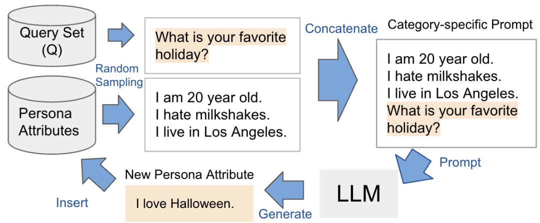



Figure 4: Query-based Persona Bootstrapping Process

$Q$ to generate new persona attributes through a bootstrapping process. We initialize $\Pi$ to $\Pi_0$. At every iteration, we randomly select a subset of persona attributes from $\Pi$, and create a set of prompts as follows: we first concatenate a set of persona attributes $s$. For every query $q \in Q$, we then combine the concatenated samples $s$, and the query $q$ to create a category-specific persona prompt. This prompt guides the LLM to generate a persona attribute for that persona category. The set of prompts obtained from this process is $\{sq|q \in Q\}$. We only add a new persona attribute to the set if its BERT embeddings (Devlin et al., 2019) are not too close from existing ones, so as to prevent the addition of duplicates.

Each of these prompts is then fed to the LLM to create a new persona attribute, which is subsequently added to the set of persona attributes $\Pi$ for the next iteration. We continue this iterative process until we have generated a total of 5k persona attributes. Figure 4 illustrates the persona bootstrapping process. Table 7 in the appendix contains the prompt template used in this component.

### 3.1.2 User Profile Construction

We build user profiles incrementally by sampling persona attributes from $\Pi_e$, and adding the eligible ones. A persona attribute is eligible if it adheres to the criteria of consistency and non-redundancy. In other words, it should not contradict any attribute already in the user profile, and it should not be inferred by other persona attribute. We assess the consistency and redundancy of user profiles by leveraging an NLI model, and persona attribute clustering, respectively. The NLI model we employ is based on T5 (Raffel et al., 2019), and has been trained on the TRUE dataset (Honovich et al., 2022).

We create a user profile $U$ by iteratively selecting a random candidate persona attribute $\pi' \in \Pi_e$. We use the NLI model to assess whether $\pi'$ contradicts any persona attribute in the profile. This is determined by the condition: $\forall \pi \in U : (\pi' \nrightarrow \neg\pi) \wedge (\pi \nrightarrow \neg\pi')$, where $\rightarrow$ is an inference. Additionally, we evaluate the similarity of $\pi'$ to the persona attributes in $U$ to prevent the addition of redundant attributes. We add $\pi'$ to $U$ if it meets the consistency and non-redundancy criteria. We repeat this process until the user profile contains 5 persona attributes. Please refer to Appendix A.1 for more details on the user profile construction.

### 3.2 User Pairing

In this component, we identify potential pairs of users for conversations. As the conversations are persona-based, we hypothesize that they will be more engaging if the users' personas exhibit more commonalities. We assign a similarity score to every pair of user profiles $(U_1, U_2)$, indicating their semantic similarity. We leverage BERT to represent the user profiles. The similarity between $U_1$ and $U_2$ is defined as: $|\{(\pi_1, \pi_2)|\pi_1 \in U_1, \pi_2 \in U_2, \exists c : \pi_1, \pi_2 \in c\}|$ Where $c$ is a persona attributes cluster. The semantic similarity is quantified by the number of common persona categories in the user profiles. We pair $U_1$ and $U_2$ if their similarity exceeds a threshold of 2.

### 3.3 Conversation Generation

Our Conversation Generation component is similar to a general-purpose dataset generation framework that generates data samples, and refines them based on a set of predefined criteria, which we refer to as *policies* (Madaan et al., 2023). The flexibility in the choice of policies for data generation allows us to emphasize different objectives. Once the active policies are selected, this component generates new data samples using a few input samples. The input to our Conversation Generation framework consists of a set of paired user profiles, a few samples of user profiles along with a persona-based conversation between them, and conversation quality metrics as policies. We follow a Generator-Critic architecture, and iteratively create the dataset following the steps shown in Figure 5:

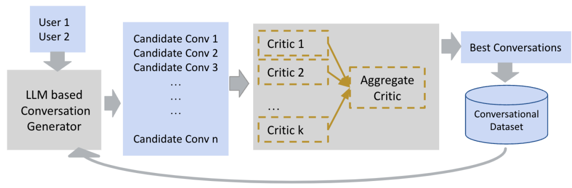



Figure 5: The Generator-Critic Architecture for Conversation Generation

**Step 1** The Generator outputs candidate conversations between persona pairs using a few initial conversation samples.
**Step 2** The Critic evaluates the candidate conversations based on the predetermined policies, and selects the best candidate conversations.
**Step 3** The best candidate conversations are added to the dataset for the next iteration of generation.
This iterative process of selecting the top candidates and adding them to the dataset gradually improves the performance of the Generator.

Without any loss of generality, we implement both the Generator and the Critic based on LLMs. Specifically, the Generator prompts an LLM to create candidate conversations, while the Critic prompts an LLM to evaluate the quality of the generated conversations.

We provide more details on the Generator, Critic, and the policies we used.

The **Generator** outputs conversations for pairs of users $(U_1, U_2)$ by prompting an LLM (Brown et al., 2020a; Wei et al., 2023). At each iteration, it randomly selects 5 samples from an initial set of conversations, each containing a pair of user profiles and a dialogue among them. It feeds these samples to a template that instructs the LLM to generate a series of candidate conversations for the given user pair. The template, and a sample generated conversation are available in Table 7, and Table 9 in the appendix.

The **Critic** selects the best generated conversations to fine-tune the Generator. A conversation is deemed high-quality if it complies with the policies of the Critic. Given the multifaceted nature of the conversation evaluations, we use a Mixture of Experts (MoE) approach. Each expert evaluates the conversation based on a specific policy. In this paper, we incorporate three types of experts, each with distinct criteria: general conversation quality, persona faithfulness, and toxicity. Collectively, these experts select the best generated conversations (the single best in our experiments). We describe each type of expert, and the collective decision-making process below.

**General Conversation Quality** experts assess conversation quality using the **Fine-grained Evaluation of Dialog (FED)** metrics introduced in (Mehri and Eskénazi, 2020). These experts use verbalized forms of the policies from FED as prompts. For instance, the "conversation depth quality expert" transforms the "depth policy" from FED into a prompt like "Which conversation is a deeper conversation between user 1 and user 2?". Our system instructs the LLM to compare each pair of candidate conversations based on these policies, resulting in pairwise comparisons. The list of policies and their baseline performance are presented in Table 6 in Appendix A.2.

The **Faithfulness** expert ensures the consistency of the generated conversations with the user profiles. It uses an LLM to identify instances of unfaithful conversations. The faithfulness prompt provides the LLM with explicit instructions, user profiles, and human-curated examples of unfaithful conversations.

The **Toxicity** expert detects any conversation that exhibits harmful traits, including bias and hate.

The Critic filters unfaithful and toxic conversations out. It then selects the best conversations using a majority vote among the General Conversation Quality experts. The selected instances are added to the dataset for the next iteration of the Generator.

# 4 Evaluation

We evaluate different aspects of our dataset generation framework, and the resulting dataset - referred to as Synthetic-Persona-Chat - which is created using an instruction fine-tuned LLM with 24 billion parameters (Chung et al., 2022). We compare Synthetic-Persona-Chat (SPC) against the widely used Persona-Chat (PC) dataset across different dimensions. We begin by evaluating the quality of the personas we generate. We then evaluate SPC using both automatic metrics, and human assessment. We analyze other aspects of SPC, such as toxicity and diversity in appendices B.1 and B.1.

## 4.1 Evaluation of the Expanded Personas

We evaluate our persona expansion module on two seed datasets: Wikipedia, and Persona-Chat. The Wikipedia personas are created by crawling the 1,000 most active contributors[2], and extracting user boxes from their pages. We expand both datasets using our framework, and evaluate the expanded persona attribute sets using automatic metrics. Table 1 compares the original persona sets to the expanded ones on a few dimensions. We observe that our persona expansion increases the number of persona attributes in SPC by 119%, while maintaining the original persona categories and expanding them by 71% compared to the persona attributes in PC. Moreover, the lengths of the new generated persona attributes are 107% longer in SPC, indicating that the new personas exhibit greater detail and specificity. We observe a similar trend when applying our persona expansion to the Wikipedia persona set, with a 108% increase in the number of persona attributes, a 140% increase in persona categories, and a 45% growth in persona attribute lengths. This demonstrates the effectiveness of our method in expanding and diversifying persona sets.

| Dataset | Persona-Chat | Synthetic-Persona-Chat | Wikipedia | Wikipedia+ |
|---|---|---|---|---|
| # Persona Attributes | 4,723 | **10,371** | 8768 | **18,293** |
| # Clusters | 323 | **553** | 408 | **986** |
| Inter-cluster Dist | 0.836 | 0.863 | 0.816 | 0.85 |
| AVG length | 7.65 | **15.9*** | 10.45 | **15.2*** |

Table 1: Evaluation of the expanded persona sets. The numbers with * indicate the metric value of the newly generated persona attributes to contrast with the initial set.

## 4.2 Next Utterance Prediction

A persona-based conversation reflects the speaker's persona explicitly or implicitly. Therefore, we expect the inclusion of information about speaker personas to enhance the performance of next utterance prediction models in such conversations. In this experiment, we assess the impact of incorporating speaker personas as prior information on both ranking, and generative - Transformer based (Vaswani et al., 2017) - next utterance prediction models. We create a subset of SPC containing conversations among user pairs included in PC for a fair comparison, i.e., for each sample in PC we have a parallel sample in SPC which has the same user pairs but different conversation between them. To create next utterance candidates, we follow PC strategy: for each utterance in a conversation in SPC, we select 19 random utterances from other conversations in the dataset. The number of train, validation and test samples in both cases are 8887, 995, 959.

[2] https://en.wikipedia.org/wiki/Wikipedia:List_of_Wikipedians_by_number_of_edits

| Method | Metric | Persona-Chat None | Persona-Chat Persona | Persona-Chat % Change | Synthetic-Persona-Chat None | Synthetic-Persona-Chat Persona | Synthetic-Persona-Chat % Change |
|---|---|---|---|---|---|---|---|
| IR Baseline | hit@1 | 18.69 | 36.86 | +97 | 19.37 (19.92) | **39.6** (26.23) | **+104** (+31) |
| Transformer (Ranker) | hit@1 | 14.24 | **19.21** | **+35** | 9.71 (64.24) | 11.74 (68.82) | +21 (+7) |
| Transformer (Generator) | hit@1 | **8.54** | 6.78 | -20 | 6.89 (41.32) | 6.66 (37.35) | **-3** (-9) |
| | Perplexity | **122.5** | 173.3 | +41 | 1032 (5.24) | 1126 (5.73) | **+9** (+9) |
| | BLUE | **0.120** | 0.094 | -21 | 0.097 (0.289) | 0.083 (0.251) | **-14** (-13) |
| | ROUGE | **0.141** | 0.113 | -24 | 0.123 (0.348) | 0.107 (0.309) | **-13** (-11) |

Table 2: Results of the next utterance prediction experiment. Performance of the trained model on the test split of PC is represented by the numbers in the table, while the numbers in parentheses indicate results for the test split of SPC.

We observe (Table 2) that the performance of ranking models increases when personas are given to the models as input for both datasets. Specifically, the Transformer (Ranker) model, known for its ability to capture conversational complexity, exhibits higher performance in SPC when evaluated on the SPC test set compared to the PC test set. However, it demonstrates relatively weaker performance when trained on the PC. This implies that SPC contains more intricate and coherent conversations.

The Transformer (Ranker) trained on SPC achieves a hit@1 of 64.24 on SPC test, 350% higher than PC (14.24). This suggests that the Transformer model can more accurately predict the next utterance in SPC, pointing to a greater coherency in conversations.

The performance of the Information Retrieval (IR) Baseline model is slightly higher for SPC: it rises by 31% when conditioned on user personas, which is lower than 97% improvement in PC. A key contributing factor for the performance improvement of the retrieval-based model (IR Baseline) on PC given the personas, is the participants' tendency to copy persona words in the conversations, whereas in SPC the personas are more implicitly reflected in the conversations. The implicit reflection of personas in SPC, makes the task more challenging for word based retrieval models, necessitating reasoning that goes beyond word level. However, when the model is trained on SPC and tested on PC, the improvement is as high as when the model is trained on PC, i.e. 104% compared to 97%.

The performance of generative models is low for this task since these models are not trained with the ranking objective. However, the performance difference while the models are conditioned on personas is lower for the model trained on SPC, with a 20% drop for the model trained on PC against 3% drop in the model trained on SPC. The increase in perplexity is 9% in SPC compared to 41% in PC. The lower rate of perplexity increase and performance drop of the model given user personas as input highlights the higher alignment of conversations with personas in SPC.

We also evaluate the performance of the next utterance prediction models when given no user, one user, and both user personas. The results suggest a higher degree of bidirectionality in SPC. We refer the reader to the Appendix B.1 for more details.

### 4.3 Human Evaluation

We compare the quality of the conversations generated by our framework against those in Persona-Chat. We randomly select 200 conversations from PC, together with their corresponding user pairs, and use our method to generate conversations among the same users. We start by following (Gehrmann et al., 2019) in running a human experiment to try and detect AI-generated content. We conduct an AI detection test where we present pairs of conversations to humans, and ask them to identify the synthetically generated one. This test is carried out on the generated conversations at the end of each iteration of creating SPC. We repeat the test for conversations generated for new persona pairs, which we refer to as iteration 3*, i.e. we pair each of these conversations with a random conversation from PC. For a robust evaluation, every pair of conversations is annotated by 3 human evaluators, and the majority vote is used as the final annotation. Details of this test are available in Appendix B.2. The results of this experiment can be found in Table 3. We observe that the losing rate of SPC is reduced by 48% from SPC Iter 1 to SPC Iter 3, and dropped below the rate of 10%. Interestingly, 91% of the conversations in SPC, which are synthetically generated, are judged as human-like as the conversations generated by humans. Moreover, conversations generated for new personas (Iteration 3*) are deemed artificial in only 8.04% of cases, showing that SPC is more realistic than PC. We also observe that in Iter 2, from 200 conversations, 79 were different from the

| Conversation Source | Lose | Win | Tie | Faithful |
|---|---|---|---|---|
| SPC Iter 1 | 17.2 | 30.1 | 52.68 | 78.5 |
| SPC Iter 2 | 18.5 | **49** | 32.5 | 80.5 |
| SPC Iter 3 | **8.8** | 35.23 | **55.95** | 76.6 |
| SPC Iter 3* | 8.04 | 32.66 | 59.29 | N/A |
| SPC (LLM2) | 11.5 | 39 | 49.5 | N/A |

Table 3: An AI detection test on 200 Generated Conversations per Iteration: Synthetic-Persona-Chat Outcomes Against Persona-Chat.

conversations in Iter 1. And in Iter 3, 57 conversations were changed compared to Iter 2. These observations suggest a decreasing rate of updates with subsequent iterations, aligning with our expectations that improvements will reach human-level conversation quality.

We also evaluate the faithfulness of the generated conversations. For each conversation, we provide annotators with a faithfulness annotation task including the speakers' persona attributes and distractor persona attribute options as shown in Figure 8. We evaluate faithfulness during 3 iterations of conversation generation for the selected 200 user pairs, and the annotators evaluate the generated conversations for each pair in every iteration. The results show that, while improving the Turing test results, faithfulness of conversations are consistently higher than 75% with at most 3% variation in between iterations, indicating high faithfulness in all iterations.

Finally, we assess the impact of LLM size on the quality of the generated dataset within our framework. We create a variant of SPC using an LLM with 540 billion parameters (LLM2). Table 3 presents human evaluations comparing the smaller LLM in multiple iterations to a single-iteration approach with LLM2. The larger model exhibits a 5% advantage in the Turing test over the first iteration of dataset generation over the smaller model. After two iterations, however, the multi-iteration approach outperforms the first iteration of the bigger model, showing our framework's capacity for cost-effective, high-quality conversation generation.

## 5 Related Work

Large Language Models (LLMs) have been used for data augmentation (Shin et al., 2021), generation (Kim et al., 2023; Dong et al., 2023; Kim et al., 2022), and evaluation (Zhang et al., 2019; Liu et al., 2023). One of the earliest works in this area (Anaby-Tavor et al., 2019) used LLMs to create a large text dataset from a small, labeled one. This idea was followed by (Wang et al., 2021; Schick and Schütze, 2021) which leveraged LLMs to create datasets without any human data. (Kumar et al., 2020) evaluated the performance of different LLMs on the data augmentation task. Several conversational dataset generation methods focused on the structure of the conversational data (Dai et al., 2022; Leszczynski et al., 2023; Abbasiantaeb et al., 2023). (Mehri et al., 2022) illustrated how LLMs can effectively generate synthetic training data for task-oriented dialogue models.

Persona-based conversations have been a popular research topic in NLP (Liu et al., 2022). One of the earliest works in this area is Persona-Chat, by (Zhang et al., 2018), which proposed the Persona-Chat dataset and evaluation metrics that have become a benchmark for persona-based conversation generation (Mazaré et al., 2018). Many subsequent works have used this dataset to train and evaluate (Mohapatra et al., 2021) their models, including DialoGPT (Zhang et al., 2020), BlenderBot (Shuster et al., 2022), and PersonaChatGen (Lee et al., 2022). PersonaChatGen automated the process of creating persona based conversations of Persona-Chat using LLMs. A challenge in generating synthetic datasets is to ensure the quality of the conversation including data faithfulness, fidelity, diversity, and consistency (Li et al., 2016; Lee et al., 2023; Veselovsky et al., 2023; Zhuo et al., 2023; Wang et al., 2023a; Mündler et al., 2023). Several works have focused on creating and using high quality training datasets (Welleck et al., 2019), and creating quality filtering components to their conversation dataset generation (Lewkowycz et al., 2022). Evaluation of the resulting conversational datasets is also challenging (Xu et al., 2021). (Wang et al., 2023b) recently introduced the paradigm of interactive evaluation of conversations with LLMs.

## 6 Conclusion and Future Work

We developed a novel framework for generating high-quality persona-based conversations using LLMs, resulting in the creation of Synthetic-Persona-Chat, comprising 20k conversations. We hope this dataset will support future endeavors in developing persona-aware conversational agents, including the generation of domain-specific multi-session conversations for specialized, task-oriented interactions. While we focused on a persona-based dataset generation task, our Generator-Critic approach can be generalized to other use cases, such as generating other specialized datasets, etc.

## Limitations

In this paper, we define an iterative process over LLMs to generate a dataset. Our method requires computational resources, and access to an LLM. The quality of the dataset is bounded by the LLM, since the quality critics are also using the same LLM, and we leave the iterative improvement of our critics as future work. The main limitation of this data generation framework is the inability to generate realistic conversations that do not have high quality, since we assume that both parties are fluent, that the conversation flow is perfectly consistent, and there is no unexpected event (e.g. an interruption by another person, connection loss, etc.) in the middle of the conversation. Another limitation of our method is the difficulty of incorporating less tangible persona traits, such as a sense of humor, or user attributes that require multiple conversation sessions to be reflected.

## Ethics Statement

The approach of generating datasets based on some desired objective might be used to create harmful datasets, and train malicious models based on them, such as a biased dataset, or a hateful speech one (Hartvigsen et al., 2022). On the other hand, these datasets and models can be used as filters in application tasks.

We used Amazon Mechanical Turk in our human experiments, and followed that platform's guidelines to protect the rights of human raters. The participation was voluntary, and the raters were informed of their rights at the beginning of the study. The platform implemented security measures to protect them, and prevent the disclosure of any Personal Identifiable Information about them. Furthermore, we offered higher than minimum standard wage compensation to avoid any exploitative practices.

To avoid having any toxic conversation in the final dataset, we also used several tools to remove any potentially toxic conversation. Details about these tools, and example removed samples are available in Appendix B.1.

## References


Zahra Abbasiantaeb, Yifei Yuan, E. Kanoulas, and Mohammad Aliannejadi. 2023. Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, N. Tepper, and Naama Zwerdling. 2019. Not enough data? deep learning to the rescue! *ArXiv*, abs/1911.03118.

Parikshit Bansal and Amit Sharma. 2023. Large language models as annotators: Enhancing generalization of nlp models at minimal cost. *ArXiv*, abs/2306.15766.

D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum. 2004. Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners.

Cheng-Han Chiang and Hung yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Annual Meeting of the Association for Computational Linguistics*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Lisbeth Claus. 2019. Hr disruption—time already to reinvent talent management. *BRQ Business Research Quarterly*, 22.

Alan Cooper and Paul Saffo. 1999. *The Inmates Are Running the Asylum*. Macmillan Publishing Co., Inc., USA.

Nico Daheim, Nouha Dziri, Mrinmaya Sachan, Iryna Gurevych, and Edoardo M. Ponti. 2023. Elastic weight removal for faithful and abstractive dialogue generation.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. *ArXiv*, abs/2205.09073.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and T. Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *ArXiv*, abs/2304.06767.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *ArXiv*, abs/2302.04166.

Kristin Fuglerud, Trenton Schulz, Astri Janson, and Anne Moen. 2020. *Co-creating Persona Scenarios with Diverse Users Enriching Inclusive Design*, pages 48–59.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. 2019. Gltr: Statistical detection and visualization of generated text. In *Annual Meeting of the Association for Computational Linguistics*.

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*, abs/2203.09509.

Xingwei He, Zheng-Wen Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2023. Annollm: Making large language models to be better crowdsourced annotators. *ArXiv*, abs/2303.16854.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Y. Matias. 2022. True: Re-evaluating factual consistency evaluation. In *Workshop on Document-grounded Dialogue and Conversational Question Answering*.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. Soda: Million-scale dialogue distillation with social commonsense contextualization.

Minju Kim, Chaehyeong Kim, Yongho Song, Seung won Hwang, and Jinyoung Yeo. 2022. Botstalk: Machine-sourced framework for automatic curation of large-scale multi-skill dialogue datasets.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. *ArXiv*, abs/2003.02245.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomás Kociský, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the gap: Assessing temporal generalization in neural language models. In *Neural Information Processing Systems*.

Dong-Ho Lee, Jay Pujara, Mohit Sewak, Ryen W White, and Sujay Kumar Jauhar. 2023. Making large language models better data creators. In *The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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.

Megan Leszczynski, Ravi Ganti, Shu Zhang, Krisztian Balog, Filip Radlinski, Fernando Pereira, and Arun Tejasvi Chaganty. 2023. Generating synthetic data for conversational music recommendation using random walks and language models. *ArXiv*, abs/2301.11489.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and William B. Dolan. 2016. A persona-based neural conversation model. *ArXiv*, abs/1603.06155.

Yen-Ting Lin and Yun-Nung (Vivian) Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *ArXiv*, abs/2305.13711.

Junfeng Liu, Christopher T. Symons, and Ranga Raju Vatsavai. 2022. Persona-based conversational ai: State of the art and challenges. *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 993–1001.

Yang Liu, Dan Iter, Yichong Xu, Shuo Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *ArXiv*, abs/2303.16634.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback.

Philip M Massey, Shawn C Chiang, Meredith Rose, Regan M Murray, Madeline Rockett, Elikem Togo, Ann C Klassen, Jennifer A Manganello, and Amy E Leader. 2021. Development of personas to communicate narrative-based information about the hpv vaccine on twitter. front digit health.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

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.

Shikib Mehri and Maxine Eskénazi. 2020. Unsupervised evaluation of interactive dialog with dialogpt. In *SIGDIAL Conferences*.

A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. 2017. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*.

Biswesh Mohapatra, Gaurav Pandey, Danish Contractor, and Sachindra Joshi. 2021. Simulated chats for building dialog systems: Learning to generate conversations from instructions.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin T. Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *ArXiv*, abs/2305.15852.

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 E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

John Pruitt and Jonathan Grudin. 2003a. Personas: Practice and theory. In *Proceedings of the 2003 Conference on Designing for User Experiences*, DUX '03, page 1–15, New York, NY, USA. Association for Computing Machinery.

John S. Pruitt and Jonathan T. Grudin. 2003b. Personas: practice and theory. In *Conference on Designing for User eXperiences*.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. *ArXiv*, abs/2104.07540.

Richard Shin, Christopher Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. 2021. Constrained language models yield few-shot semantic parsers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7699–7715, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael Shum, Stephan Zheng, Wojciech Kryscinski, Caiming Xiong, and Richard Socher. 2019. Sketch-fill-a-r: A persona-grounded chit-chat generation framework. *ArXiv*, abs/1910.13008.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, W.K.F. Ngan, Spencer Poff, Naman Goyal, Arthur D. Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *ArXiv*, abs/2208.03188.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *ArXiv*, abs/1409.3215.

Michelle van Pinxteren, Mark Pluymaekers, and Jos Lemmink. 2020. Human-like communication in conversational agents: a literature review and research agenda. *Journal of Service Management*, ahead-of-print.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023. Generating faithful synthetic data with large language models: A case study in computational social science.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zi-Han Lin, Yuk-Kit Cheng, Sanmi Koyejo, Dawn Xiaodong Song, and Bo Li. 2023a. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *ArXiv*, abs/2306.11698.

Xiaolei Wang, Xinyu Tang, Wayne Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023b. Rethinking the evaluation for conversational recommendation in the era of large language models.

Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. 2021. Towards zero-label language learning. *ArXiv*, abs/2109.09193.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference.

Chien-Sheng Wu, Andrea Madotto, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2019. Getting to know you: User attribute extraction from dialogues. In *International Conference on Language Resources and Evaluation*.

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

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur D. Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Annual Meeting of the Association for Computational Linguistics*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation.

Peixiang Zhong, Yao Sun, Yong Liu, Chen Zhang, Hao Wang, Zaiqing Nie, and Chunyan Miao. 2020. Endowing empathetic dialogue systems with personas. *ArXiv*, abs/2004.12316.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of XiaoIce, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity.

# A Dataset Generation Framework

In this section, we provide more details on our synthetic dataset generation framework. We created Synthetic-Persona-Chat using an LLM with 24 billion parameters. We use top-k sampling with $k = 40$ for decoding during generation, and set the temperature value to 0.7 in all components. We give more details on user and conversation generation components in the following subsections.

## A.1 User Generation

In our framework, the user generation component consists of two steps: expanding the persona attribute set, and creating realistic user profiles. In this section we provide details on our framework for these two steps:

**Persona Expansion** As described in Section 3.1.1, the persona expansion step involves identifying persona categories in the initial persona attribute set $\Pi_0$, generating queries associated with those categories, and bootstrapping queries to create a query set . In our framework, we employ the Scikit-learn (Pedregosa et al., 2011) implementation of an agglomerative clustering to identify persona categories following this clustering method: we represent each persona using a BERT-based representation. Our clustering approach is bottom-up, starting with each persona attribute as an individual cluster. At each step, we combine two clusters if their similarity exceeds a predetermined threshold of 0.1. The similarity of two clusters is measured using inter-cluster average cosine similarity. The process continues until no pair of clusters is more similar than the threshold. We set the value of the threshold as 0.1 since it lead to more than 100 non-sparse clusters, i.e., clusters that include at least 3 persona attributes and can be used in the query induction prompt. Table 4 presents the cluster similarity threshold values and the resulting cluster details based on them.

| Similarity Threshold | # Clusters | # Sparse Clusters |
|---|---|---|
| 0.05 | 1083 | 171 |
| 0.1 | 323 | 6 |
| 0.15 | 17 | 2 |

Table 4: Details of persona clusters created based on similarity threshold in agglomerative clustering.

After identifying the clusters, we sample 3 instances of persona attributes for each cluster, and prompt the LLM using the template in shown in section 3 to construct an initial query set $Q_0$. We expand the query set $Q_0$ using bootstrapping. At each step, we sample 5 instances from the available queries, and prompt the LLM using the template in Table 7. We repeat this process for 100 steps. Examples of initial persona attributes, induced queries, bootstrapped queries, and bootstrapped persona attributes can be found in Table 5. The prompt templates used in this component are available in Table 7.

**User Profile Generation** We illustrate a sample user profile creation process in Figure 6. As shown in the figure, at each iteration, a randomly selected persona attribute is checked for consistency and non-redundancy.

Let $\pi'$ be a randomly selected persona attribute in an iteration. For the redundancy criteria, we use the BERT representation of persona attributes. We compute the similarity of the new candidate persona attribute $\pi'$ with every persona attribute in the user profile. If it is more than a threshold (0.9 in these experiments) similar to an attribute in the user profile, $\pi'$ is deemed as redundant and will not be added to the user profile. We use the cosine similarities of the BERT representations of the persona attributes. The value of the similarity threshold is selected to be compatible with the agglomerative persona clustering algorithm in the persona expansion step, in which two clusters are merged if their inter-distance is less than 0.1, i.e., their inter-cluster similarity is higher than 0.9. Therefore, by setting the threshold of similarity of attributes to be .9, we ensure that the new attribute is added to the user profile if it is from a new cluster compared to the current attributes in the user profile.

For the consistency criteria, we use the NLI model to verify the consistency of this persona attribute with the user profile. For every persona attribute in the current user profile $\pi$, we prompt the LLM to create the negated persona attribute $\neg\pi$. Then, we query the NLI model to check whether $\neg\pi$ is inferred by $\pi'$ or $\neg\pi'$ is inferred by $\pi$. If either of these cases is inferred, then the selected persona attribute is not consistent with the user profile, and not added to the profile.

## A.2 Conversation Generation

**LLM-based Critic** In our framework, the critic is implemented by prompting an LLM. We included a mixture of experts approach in the critic, where each expert prompts the LLM to assess a

| Dataset | Persona Source | Query | Example Persona Attribute |
|---|---|---|---|
| Persona-Chat | Human | What is your job?<br>Where do you live?<br>Do you have any pets? | I am a pharmacist.<br>I live close to the coast.<br>I have a doberman. |
| | LLM | What are your talents?<br>What is your hair color?<br>What is your favorite song? | I am a great listener.<br>My hair is auburn.<br>I like the song "Leather and Lace". |
| Wikipedia | Human | What are your hobbies?<br>What is your view on the metric system? | I spend WAY too much time on Wikipedia.<br>I find the metric system to be a logical and efficient way to measure things. |
| | LLM | What is the name of the first album you ever purchased?<br>What are you interested in? | My first album was The Miseducation of Lauryn Hill<br>I'm looking to learn new recipes and improve my cooking skills. |

Table 5: Persona Categories and Induced Queries Using Our Framework. Queries are generated by the Large Language Model (LLM). Queries for personas with the "LLM" as source, are generated through bootstrapping, while those with "human" as source are generated by sampling persona categories and prompting the LLM. Personas with "human" as the source are authored by humans, while "LLM" rows represent personas generated using our framework.


Figure 6: User Profile Construction Example

| Policy | Performance |
|---|---|
| Depth | 0.84 |
| Coherency | 0.96 |
| Consistency | 0.92 |
| Diversity | 0.92 |
| Likable | 0.88 |

Table 6: List of FED Experts for Persona-Based Conversation Generation Critic. Performance is measured by the number of correctly compared conversation pairs in FED baseline based on the given policy.

specific policy in the candidate conversations. Our framework includes a set of experts to control the general conversation quality. We evaluate the performance of these experts using a baseline dataset. The baseline dataset for this experiment is FED which consists of 125 human-annotated instances evaluated at the conversation level. We pair the conversations and evaluate the experts based on the number of correctly ranked pairs. As shown in Table 6, we observe that these experts are more than 80% accurate in distinguishing the better conversation within the pairs. The template for the verbalized form of these experts used in our framework can be found in Table 7.

We also included a toxicity expert and a persona faithfulness expert in the critic. The prompt templates used in these experts are available in Table 7. The persona faithfulness leverages in-context-learning capability of LLMs. It includes a few human-curated examples of faithful and unfaithful conversations in the instruction prompt. Refer to Table 8 for examples of faithful and unfaithful conversations used in the instruction prompt.

The faithfulness critic, prompts the LLM both with and without the candidate conversation between two users. It assesses the log probabilities of the output being "Yes" (indicating a contradiction and thus unfaithfulness) or "No" (indicating no contradiction and thus faithfulness). A conversation is deemed unfaithful if there is an increase in the

| Component | Template |
|---|---|
| Query Induction | What is the most specific question that you are replying to with the following statements?<br>{persona-category-sample-1}<br>{persona-category-sample-2}<br>{persona-category-sample-3} |
| Query Bootstrapping | {cluster-query-1}<br>...<br>{cluster-query-5}<br>Add more persona questions similar to the above examples. |
| Persona Bootstrapping | Imagine you are a person with the following persona.<br>{random-persona-attribute-1}<br>...<br>{random-persona-attribute-5}<br>{query}. Answer with only one short sentence that starts with 'I' or 'My'. Do not repeat the given persona. |
| FED Expert | Which one of Conversation 1 and Conversation 2 between two users {policy}? Why?<br>Conversation 1: {conv-1}<br>Conversation 2: {conv-2} |
| Toxicity Expert | Is this conversation toxic? Why?<br>Conversation: {conv} |
| Conversation Generation | Here, we list the profiles of two users, user 1 and user 2, followed by an interesting and natural conversation between user 1 and user 2, which implicitly reflects their user profiles.<br>User 1 Profile: {conversation1-user-1}<br>User 2 Profile: {conversation1-user-2}<br>Conversation: {conversation-1}<br>...<br>User 1 Profile: {conversation-5-user-1}<br>User 2 Profile: {conversation-5-user-2}<br>Conversation: {conversation-5}<br>Give me more examples like this. The conversation must be more than 5 turns and less than 8 turns. The conversation must be natural, and not direct copies of their profiles.<br>User 1 Profile: {user-1}<br>User 2 Profile: {user-2} |
| Faithfulness Expert | Given user 1 and user 2's profiles respectively, does the following conversation between the two users contradict either of their profiles? Why?<br>User 1 Profile: {user-1}<br>User 2 Profile: {user-2}<br>Conversation: {conv-1}<br>Response: {explanation} |

Table 7: Prompting Templates for Large Language Models of Different Components in Our Framework. Variables enclosed in {} are filled when the template is populated.

probability of a contradiction ("Yes") or a decrease in the probability of no contradiction ("No").

## B Synthetic-Persona-Chat

Synthetic-Persona-Chat is made of 20k conversations, with an average of 11.8 turns per user for each. An example Synthetic-Persona-Chat conversation can be found in Table 9. We compare Synthetic-Persona-Chat to Persona-Chat across different dimensions. We first assess the characteristics of SPC using various automatic evaluators, i.e. evaluators which do not require human effort. We then conduct a human evaluation experiment on a subset of SPC.

### B.1 Automatic Evaluation

We conduct a comprehensive analysis and evaluation of SPC across different dimensions and compare it against PC. We start by analyzing the toxicity and diversity of SPC using off the shelf tools. Then, we elaborate on the experiments which assess the efficacy of SPC used as the dataset for the next utterance prediction and the profile extraction tasks. Finally, we evaluate the quality of SPC conversations using LLM-based evaluation methods.

**Toxicity Analysis** We analyze the toxicity of the generated conversations at the final iteration of SPC using an online tool called Perspective[3]. We reproduce the results of a detailed analysis of toxicity in PC as well as in each iteration of our data generation framework while producing SPC in Table 10. We observe a notable reduction in the frequency of conversations deemed as strongly toxic or profane throughout the iterations of generating SPC. This reduction can be attributed to the built-in toxicity filter of the employed LLM. While PC contains more than 50 samples that are identified as strongly toxic, SPC includes at most three toxic or profane conversations, which is significantly lower (at least 15 times less). Interestingly, the fraction of conversations with medium profanity and toxicity in SPC is 4 times less than the same type of conversations in PC across all iterations. We have removed any conversation that was marked as strongly toxic by this tool in the released dataset. Samples of toxic conversations are provided in Table 11.

**Diversity Analysis** We use hierarchical topic modeling (Blei et al., 2004) to assess the topic diversity of SPC and compare it to that of PC. For a fair comparison, we only compare conversations in SPC with similar personas in PC. Table 12 displays the number of topics at each level of the topic tree, with the first level indicating the most general topic. We observe similar topic diversity at the first level. In deeper levels, there is a slightly lower diversity in SPC.

**Next Utterance Prediction** We compare the performance of different models on the next utterance prediction task. As discussed in Section 4.2, these models are expected to exhibit better performance in the next utterance prediction task when user personas are provided as prior information. We evaluate ranking and generative models for response selection to assess this property. We compare models trained on SPC to the same models trained on PC. We use the implementations provided in (Miller et al., 2017) for the following models:

- **IR Baseline** Given an utterance as a query, the IR baseline finds the most similar utterance in the training corpus using tf-idf. It defines the utterance after the most similar utterance as the candidate response, and then returns the most similar option to that candidate as the output.
- **Transformer-Ranker** The context of the conversation, as well as the candidate next utterances, are encoded using a BERT-based encoder. The most similar encoded candidate to the conversation context, as measured by a dot-product in their representation space, is selected as the output (Humeau et al., 2020).
- **Transformer-Generator** This model is a sequence-to-sequence model (Sutskever et al., 2014) which uses transformers as encoders and decoders.

We also evaluate the performance of the next utterance prediction models when given no user, one user, and both user personas. The results of this experiment are available in Table 13. We observe that the highest performance improvement for all models trained on PC is when self-personas are given as input. We do not observe such a pattern in SPC. This indicates a higher degree of bidirectionality in SPC conversations compared to those of PC.

**Profile Extraction** A potential use-case of the SPC dataset is training a model to predict user personas from a conversation. This is only possible if

[3]https://perspectiveapi.com/

| User Profiles | Conversations | Faithfulness Explanation |
|---|---|---|
| **User 1**:<br>One of my favorite hobbies is gaming.<br>I currently work in an office.<br>I enjoy running<br><br>**User 2**:<br>I am applying for publishing jobs.<br>My mother is a librarian.<br>The only autographs I've ever wanted are from authors.<br>I used to wear glasses as a child.<br>My favorite color is green. | User 1: Do you like video games?<br>User 2: No, I do not like video games.<br>User 1: What do you do for a living?<br>User 2: I am a librarian.<br>User 1: I guess that is why you like to read a lot.<br>User 2: Yes, I like to read books.<br>User 1: What is your favorite color? Mine is green.<br>User 2: That is my favorite color too.<br>User 1: That is cool.<br>User 2: Do you have any pets?<br>User 1: No, I do not have any pets.<br>User 2: I have a dog and a cat.<br>User 1: That is cool.<br>User 2: Do you have any hobbies?.<br>User 1: Yes, I like to write.<br>User 2: What do you write about?<br>User 1: I write about my life.<br>User 2: That is cool.<br>User 1: Thanks! | No, the conversation does not contradict the profiles. Neither user 1 nor user 2 says anything that contradicts their respective profiles. Although the conversation does not cover everything in their profiles, that is ok. |
| **User 1:**<br>I would like to move some place warmer.<br>I m an out of work farmhand.<br>I like watching reality television shows.<br>I live in Kansas.<br><br>**User 2:**<br>I am studying to be a dentist.<br>My favorite team is the raptors.<br>My favorite band is rise against.<br>I usually go to bed late.<br>I like sports. | User 1: It's so cold in Kansas, I want to go some place warmer.<br>User 2: I am studying dentistry.<br>User 1: I am an out of work farmhand, I love reality tv.<br>User 2: I like the raptors and rise against, I go to bed late.<br>User 1: It is very hard to find work here, I might need to go to the city.<br>User 2: I don't like sports , but I like the raptors .<br>User 1: I would like to work as a model, but I have no experience.<br>User 2: I am applying for a job as a receptionist.<br>User 1: I am very jealous, I hope you get it. | Yes, the conversation contradicts user 2's profile. In the profile, user 2 says they like sports, however, in the conversation, user 2 says they don't like sports. |
| **User 1**:<br>My family is from Kazakhstan.<br>I can speak Spanish, English, and Kazakh.<br>I like to go hiking on the weekends.<br>I play jazz piano in a band.<br>I am a barista.<br><br>**User 2:**<br>I am a vegan.<br>I love game of thrones.<br>I like soccer.<br>My brother used to be in the navy. | User 1: Hi, what do you do for a living?<br>User 2: I am a barista.<br>User 1: Awesome, me too! Where are you from?<br>User 2: My family is from Kazakhstan, but I was born in Uruguay.<br>User 1: Cool . What do you like to do in your free time?<br>User 2: I love to go hiking and play jazz piano.<br>User 1: Wow, that sounds really nice. I like to watch game of thrones and play soccer.<br>User 2: I have never seen game of thrones, but I love soccer.<br>User 1: You should watch it, it is really good.<br>User 2: I will check it out. | Yes, the conversation contradicts user 2's profile. In the profile, user 2 says they love game of thrones, but in the conversation user 2 says they have never seen it. |
| **User 1:**<br>My favorite book is David Copperfield.<br>I have a twin brother.<br>I am really short.<br>I love to bake cakes.<br>I have won tap dance competitions.<br><br>**User 2:**<br>I enjoy spending a lot of money on my hobbies.<br>I love photography.<br>I like to use candles instead of lights.<br>I do not like waking up early. | User 1: Hi there! Do you like to read?<br>User 2: I like to read. I like to go to the library.<br>User 1: Oh, cool! What is your favorite book?<br>User 2: I really like David Copperfield. I have a twin brother and I am really short.<br>User 1: Are you good at baking?<br>User 2: I can bake cakes. I have won tap dance competitions.<br>User 1: Do you like photography?<br>User 2: I love taking pictures. I also like to use candles instead of lights.<br>User 1: Do you like to wake up early?<br>User 2: I like to sleep in. I do not like waking up early.<br>User 1: Do you like to spend money?<br>User 2: I love spending money on my hobbies. | No, the conversation does not contradict either user profiles. Although user 2 says they are really short, which is in user 1's profile, it doesn't mean user 2 cannot be really short too. |

Table 8: Examples of Faithful and Unfaithful Conversations for Prompting the Persona Faithfulness Expert in Our Framework. The 'Faithfulness explanation' columns are used to fill the 'explanation' variable in the faithfulness expert template.

| User Profiles | Conversation |
|---|---|
| **User 1**<br>I am male.<br>My weakness are sweets.<br>I am unsociable.<br>I am a jack of all, master of none.<br>**User 2**<br>My favorite music is rock and roll.<br>I love to travel.<br>I am from Russia.<br>I enjoy skiing.<br>I go to school for business. | User 1: Hello<br>User 2: Hello, what's your name?<br>User 1: My name is [name], and you?<br>User 2: My name is [name], I'm from Russia.<br>User 1: Nice to meet you, [name]. I'm from the United States.<br>User 2: It's nice to meet you too, [name]. What are you interested in?<br>User 1: I'm interested in a lot of things, but my main interests are music and travel.<br>User 2: I love music too! What kind of music do you like?<br>User 1: I like a lot of different kinds of music, but my favorite is rock and roll.<br>User 2: Me too! I love classic rock and roll. Who are your favorite bands?<br>User 1: My favorite bands are The Beatles, The Rolling Stones, and Led Zeppelin.<br>User 2: Those are all great bands! I love them too.<br>User 1: What about you, what are your favorite bands?<br>User 2: My favorite bands are The Doors, Pink Floyd, and Queen.<br>User 1: Those are all great bands too! I love them too.<br>User 2: So, what do you like to do for fun?<br>User 1: I like to go to concerts, play video games, and watch movies.<br>User 2: I love going to concerts too! I also like to go skiing and play tennis.<br>User 1: I love skiing too! I haven't played tennis in a while, but I used to play a lot.<br>User 2: We should go skiing together sometime!<br>User 1: That would be fun! I'm free this weekend if you are.<br>User 2: I'm free this weekend too! We should go skiing.<br>User 1: Great! I'll text you the details. |

Table 9: Sample Conversation from Synthetic-Persona-Chat. This conversation was synthesized from user profiles in Persona-Chat

| | Toxicity | | | Profanity | | |
|---|---|---|---|---|---|---|
| Confidence | weak(< .2) | medium(.2-.8) | strong(>.8) | weak(< .2) | medium(.2-.8) | strong(>.8) |
| PC | 10875 | 4448 | 53 | 10891 | 1676 | 57 |
| SPC Iter 1 | 10902 | 1192 | 3 | 10903 | 340 | 3 |
| SPC Iter 2 | 10900 | 1096 | 1 | 10901 | 345 | 1 |
| SPC Iter 3 | 10902 | 1088 | 1 | 10902 | 376 | 0 |

Table 10: Frequency of Toxic Conversations in Persona-Chat and Synthetic-Persona-Chat

| Source | Conversation |
|---|---|
| Persona-Chat | ...<br>User 1: I like bloody stuff.<br>User 2: It reminds me of the dark which makes me afraid of it.<br>User 1: You are a silly goose. |
| Persona-Chat | ...<br>User 2: Cool. Why do you say that? Because I am a red head?<br>User 1: No. Ikn. Why do you ask so many questions? Mr. Thomas is dumb. |
| Synthetic-Persona-Chat | User 1: I can imagine. What's your favorite part of the job?<br>User 2: I love working with my team and seeing our restaurant succeed.<br>User 1: That's great. What's your least favorite part of the job?<br>User2: My least favorite part is dealing with my boss. He's a real jerk. |

Table 11: Examples of Toxic Conversations. The first two examples are segments of conversations from Persona-Chat. The final example is a segment from a toxic conversation in Synthetic-Persona-Chat, which has been removed in the released dataset.

| Topic Level | PC | SPC |
|---|---|---|
| 1 | 27 | 27 |
| 2 | 232 | 213 |
| 3 | 470 | 403 |
| 4 | 137 | 118 |
| 5 | 30 | 26 |

Table 12: Vertical Topic Diversity in Persona-based Datasets

the dataset is highly faithful, meaning that any persona attribute inferred from the conversation is in the user profile or compatible with the user profile. In this context, a faithful conversation is expected to have high precision in the profile extraction task, while a conversation that highly reflects user personas is expected to have high recall in this task.

We evaluate the task of user profile extraction for conversations in SPC, and compare the results against those of PC. We frame the task of profile extraction as a ranking task, using the utterances within the conversations as queries. The goal is to rank a set of persona attribute options. For each conversation, we include the speakers' persona attributes in the available options. Additionally, we select 25 random user persona attributes from other speaker profiles within the dataset to serve as distractors. The input to the profile extraction is utterances from a single user as the speaker, while the output is a list of persona attribute options for a target user, which could be either user 1 or user 2. The results of this experiment are presented in Table 14. We observe that the performance of the profile extraction methods is higher in SPC in 3 of the 4 scenarios. Interestingly, we observe that with both datasets, when the target and the speaker are different, the performance of profile extraction is greater compared to the cases when the target and speaker users are the same.

**LLM-based Quality Evaluation** We leverage LLM-based conversation quality evaluators from the literature to compare the quality of SPC and PC. These evaluators rely on the human curated prompt templates for different metrics including consistency, fluency, etc. We used these evaluators with minimum change in the original prompt templates. These evaluators are:

- **LLM-Eval** (Lin and Chen, 2023) is a multi-dimensional automatic evaluation designed for conversations. It uses a human-curated prompt which describes evaluation dimensions, serving as a unified evaluation schema. This prompt evaluates the conversation across multiple dimensions (e.g. fluency) in a single model call. We show this unified schema in Table 15.

- **GPT-Score** (Fu et al., 2023) leverages emergent abilities of LLMs, i.e. zero-shot instructions, to score texts. It contains a prompt template, and for each quality criterion, populates the template with a human description of the criteria along with the valid score range for that criteria. Example prompts are provided in Table 15.

- **G-Eval** (Liu et al., 2023) introduces a framework that employs LLMs with a chain-of-thought approach to assess the quality of natural language generated outputs. For any evaluation criteria, G-Eval prompts the LLM with the criterion's description, prompting the

| Method | Metric | Persona-Chat No Persona | Self Persona | Their Persona | Both Personas | Synthetic-Persona-Chat No Persona | Self Persona | Their Persona | Both Personas |
|---|---|---|---|---|---|---|---|---|---|
| IR baseline | hit@1 | 0.1869 | **0.3683** | 0.1519 | 0.3281 | 0.1861 | **0.2596** | 0.1882 | 0.2493 |
| Transformer(Ranker) | hit@1 | 0.2513 | **0.275** | 0.1922 | 0.2572 | 0.7164 | 0.6227 | 0.6988 | **0.7214** |
| Transformer (Generator) | hit@1 | **0.0896** | 0.08512 | 0.0873 | 0.0813 | 0.0526 | **0.629** | 0.053 | 0.051 |
| | ppl | 65.57 | 72.24 | **62.49** | 64.07 | 5.54 | 5.47 | **5.4** | 5.405 |

Table 13: Evaluation of Next Utterance Prediction models conditioned on different user personas.

| Target | Speaker | F-Score PC | F-Score SPC |
|---|---|---|---|
| user 1 | user 1 | 0.505 | **0.574** |
| user 1 | user 2 | **0.737** | 0.68 |
| user 2 | user 1 | 0.50 | **0.57** |
| user 2 | user 2 | 0.456 | **0.494** |

Table 14: Accuracy of Profile Extraction in Four Different Scenarios. The 'Target' column represents the user profile to be extracted, while the 'Speaker' column indicates the speaker of the turns given to the model as input.

model to generate the necessary evaluation steps. It then uses these steps to prompt the LLM to score given output for that criterion. It considers the probability of getting each permissible score as the output of the prompt, i.e., it considers the probability distribution of scores assigned by the LLM. The reported output is the expected value of the score distribution by the LLM. Table 15 includes an example prompt.

Results of this evaluation are presented in Table 16. We observe that SPC consistently outperforms PC across all the dimensions we evaluate. The superiority of SPC is more prominent when using GPT-Score, for which each evaluated criterion shows an average improvement of at least 23 points.

### B.2 Human Evaluation

We run a human evaluation of the performance of our method via a crowdsourcing platform. We conduct an AI detection test, and a faithfulness study - both of which we describe in more details in the following subsections - at the end of every iteration of the generation of SPC.

**AI Detection Test** We randomly select 200 user pairs from PC. For each example, we show the annotators the user pair, together with the corresponding conversations from PC and SPC, and ask them to select the conversation that was synthetically generated. We show an example of this crowdsourcing task in Figure 7. The results of the AI detection test are available in Table 17. We report the losing rate of SPC in the AI detection test, and Fleiss' Kappa to assess the inter-rater agreement. The agreement falls into the fair to moderate agreement bucket.

**Faithfulness** We present the annotators with a conversation, and a set of options of persona attributes. The annotators are asked to select the user persona attributes they would infer from the conversation. Figure 8 shows a sample of the annotation task in this study. The options include the persona attributes of the speakers in the conversation, and a set of distractor persona attributes. We created distractor persona attributes using different strategies to cover different difficulty levels. For a persona attribute set $\Pi$, we create a set $\neg\Pi$ of distractor persona attributes as:

**Negated personas** We prompt an LLM to negate persona attributes. For example, the negation of persona attribute "I like vegetables" is "I don't like vegetables".

**Random personas** We randomly select persona attributes from user profiles in other conversations in the dataset.

**Contradicting personas** We prompt an LLM to generate a persona attribute which contradicts the users' personas.

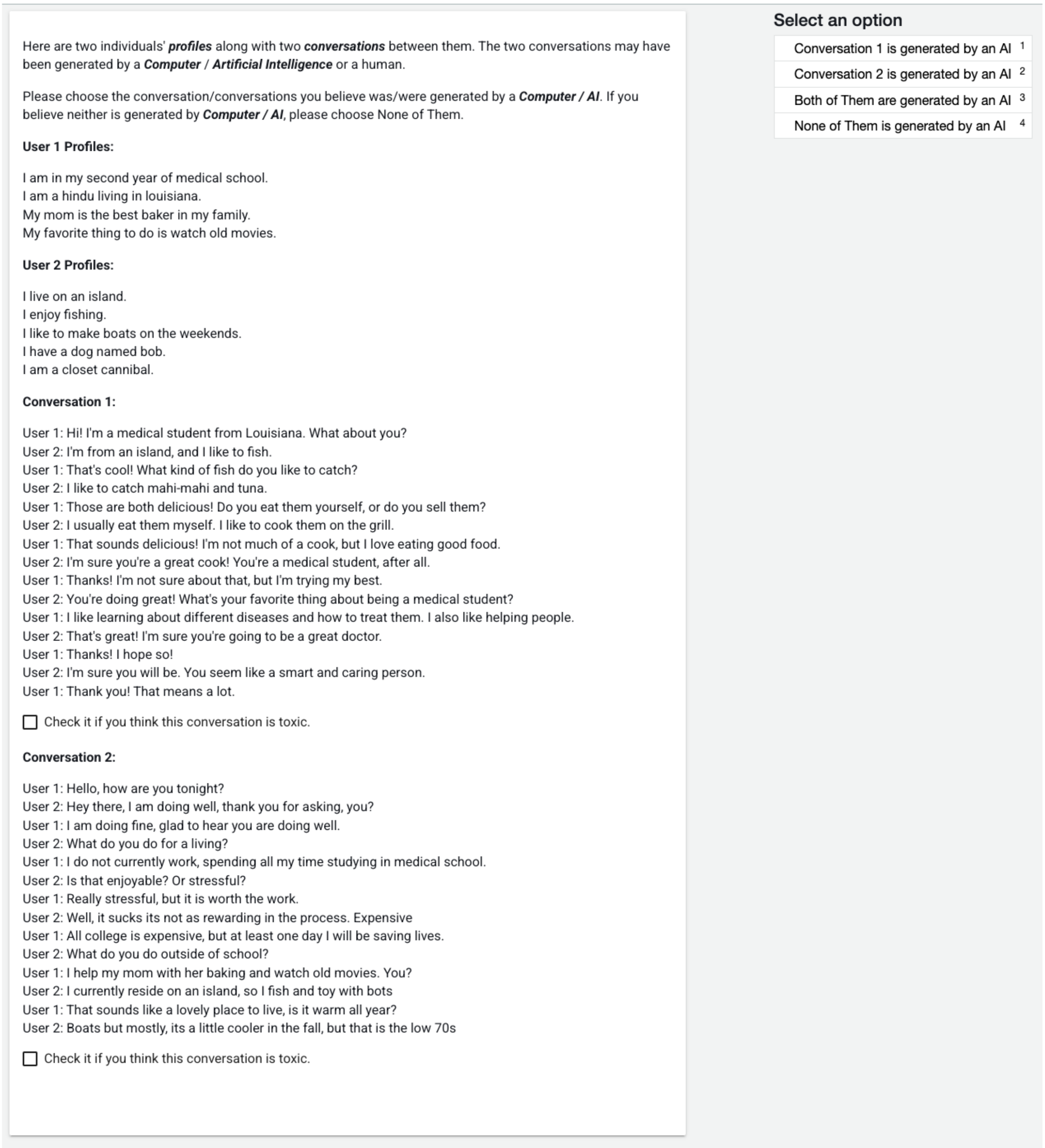
Each entry of this task includes 8 user persona attributes as options, where 4 of them are the real persona attributes, and the other 4 are distractors. We evaluate the precision of the human annotators, and report it as a proxy to the conversation faithfulness in Table 3.

## C Ablation Studies

We run several ablation studies to evaluate the importance of individual components in our framework. We begin by analyzing the effect of the persona expansion module. We then review the impact of each expert in the mixture forming our Critic.

<table>
<tr><th>Evaluator</th><th>Metric</th><th>Prompt Template</th></tr>
<tr><td>LLM-Eval</td><td>All</td><td>Human: The output should be formatted as a JSON instance that conforms to the JSON schema below.<br><br>As an example, for the schema {"properties": {"foo": {"title": "Foo", "description": "a list of strings", "type": "array", "items": {"type": "string"}}}, "required": ["foo"]}} the object {"foo": ["bar", "baz"]} is a well-formatted instance of the schema. The object {"properties": {"foo": ["bar", "baz"]}} is not well-formatted.<br><br>Here is the output schema: {"properties": {"content": {"title": "Content", "description": "content score in the range of 0 to 100", "type": "integer"}, "grammar": {"title": "Grammar", "description": "grammar score in the range of 0 to 100", "type": "integer"}, "relevance": {"title": "Relevance", "description": "relevance score in the range of 0 to 100", "type": "integer"}, "appropriateness": {"title": "Appropriateness", "description": "appropriateness score in the range of 0 to 100", "type": "integer"}}, "required": ["content", "grammar", "relevance", "appropriateness"]}<br><br>Score the following dialogue generated on a continuous scale from {score-min} to {score-max}.<br>Dialogue: {dialogue}</td></tr>
<tr><td>GPT-Score</td><td>Consistency</td><td>Answer the question based on the conversation between two users.<br>Question: Are the responses of users consistent in the information they provide throughout the conversation? (a) Yes. (b) No.<br>Conversation: {dialogue} Answer:</td></tr>
<tr><td>G-Eval</td><td>Coherence</td><td>You will be given a pair of user personas. You will then be given one conversation between this persona pair.<br>Your task is to rate the conversation on one metric.<br>Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.<br><br>Evaluation Criteria:<br><br>Coherence (1-5) - the collective quality of all utterances. We align this dimension with the Document Understanding Conference (DUC) quality question of structure and coherence , whereby "the conversation should be well-structured and well-organized. The conversation should not just be a heap of related information, but should build from utterance to a coherent body of conversation about a topic."<br><br>Evaluation Steps:<br><br>1. Read and understand the given conversation between the pair of user personas.<br>2. Evaluate the conversation based on the coherence of the utterances.<br>3. Rate the conversation on a scale of 1 to 5, with 5 being the highest coherence and 1 being the lowest coherence.<br>4. Justify the rating by referring to specific aspects of the conversation that demonstrate its coherence or lack thereof.<br><br>Example:<br><br>Personas: {personas}<br>Conversation: {dialogue}<br><br>Evaluation Form (scores ONLY):<br>- Coherence:</td></tr>
<tr><td rowspan="2">LLM-Faithfulness</td><td>Inference</td><td>Instruction: Select User {user} persona attributes that are directly inferred from this conversation.</td></tr>
<tr><td>Contradiction</td><td>Instruction: Select User {user} persona attributes that strongly contradict this conversation.</td></tr>
</table>

Table 15: Prompt Templates in LLM-based Conversation Quality Evaluators. Variables enclosed in {} are filled when the template is populated.



Here are two individuals' ***profiles*** along with two ***conversations*** between them. The two conversations may have been generated by a ***Computer*** / ***Artificial Intelligence*** or a human.

Please choose the conversation/conversations you believe was/were generated by a ***Computer / AI***. If you believe neither is generated by ***Computer / AI***, please choose None of Them.

**User 1 Profiles:**

I am in my second year of medical school.
I am a hindu living in louisiana.
My mom is the best baker in my family.
My favorite thing to do is watch old movies.

**User 2 Profiles:**

I live on an island.
I enjoy fishing.
I like to make boats on the weekends.
I have a dog named bob.
I am a closet cannibal.

**Conversation 1:**

User 1: Hi! I'm a medical student from Louisiana. What about you?
User 2: I'm from an island, and I like to fish.
User 1: That's cool! What kind of fish do you like to catch?
User 2: I like to catch mahi-mahi and tuna.
User 1: Those are both delicious! Do you eat them yourself, or do you sell them?
User 2: I usually eat them myself. I like to cook them on the grill.
User 1: That sounds delicious! I'm not much of a cook, but I love eating good food.
User 2: I'm sure you're a great cook! You're a medical student, after all.
User 1: Thanks! I'm not sure about that, but I'm trying my best.
User 2: You're doing great! What's your favorite thing about being a medical student?
User 1: I like learning about different diseases and how to treat them. I also like helping people.
User 2: That's great! I'm sure you're going to be a great doctor.
User 1: Thanks! I hope so!
User 2: I'm sure you will be. You seem like a smart and caring person.
User 1: Thank you! That means a lot.

☐ Check it if you think this conversation is toxic.

**Conversation 2:**

User 1: Hello, how are you tonight?
User 2: Hey there, I am doing well, thank you for asking, you?
User 1: I am doing fine, glad to hear you are doing well.
User 2: What do you do for a living?
User 1: I do not currently work, spending all my time studying in medical school.
User 2: Is that enjoyable? Or stressful?
User 1: Really stressful, but it is worth the work.
User 2: Well, it sucks its not as rewarding in the process. Expensive
User 1: All college is expensive, but at least one day I will be saving lives.
User 2: What do you do outside of school?
User 1: I help my mom with her baking and watch old movies. You?
User 2: I currently reside on an island, so I fish and toy with bots
User 1: That sounds like a lovely place to live, is it warm all year?
User 2: Boats but mostly, its a little cooler in the fall, but that is the low 70s

☐ Check it if you think this conversation is toxic.

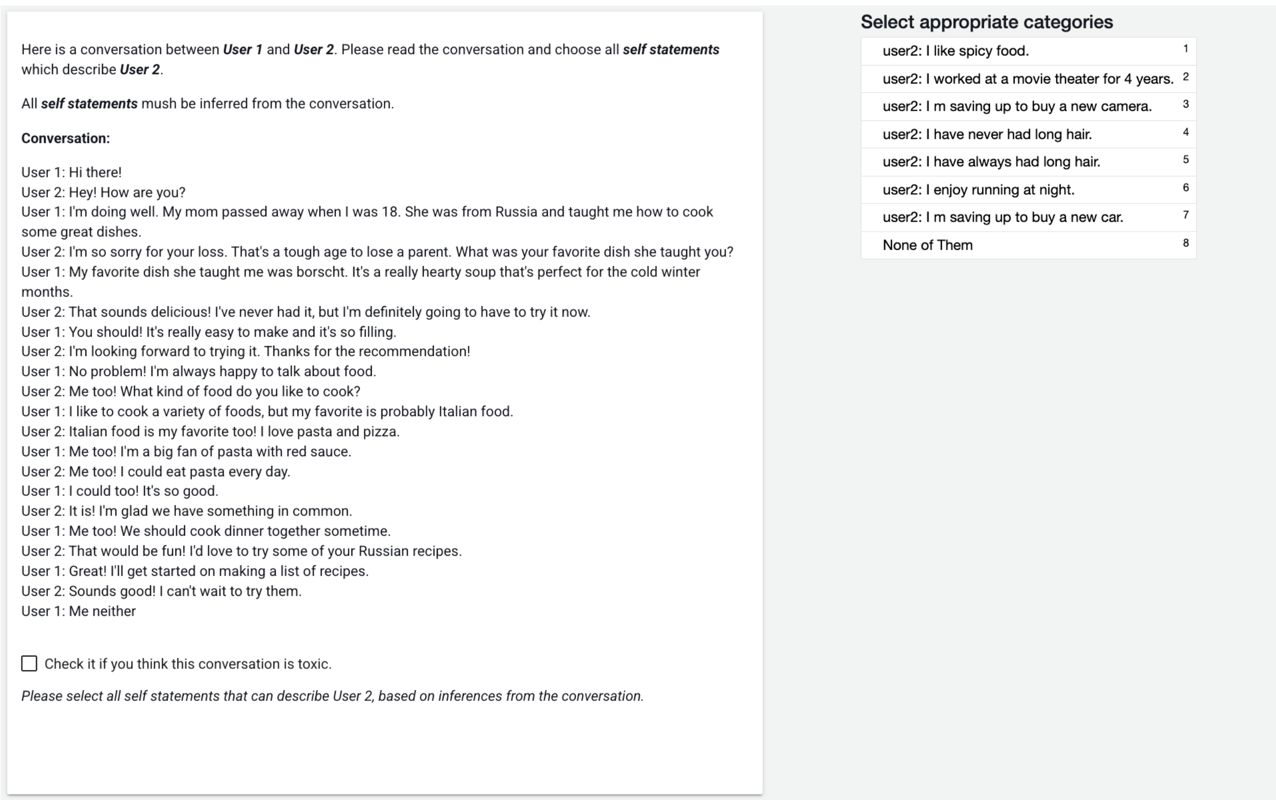
**Select an option**

- Conversation 1 is generated by an AI 1
- Conversation 2 is generated by an AI 2
- Both of Them are generated by an AI 3
- None of Them is generated by an AI 4

Figure 7: Preview of the AI detection test Task on the Crowdsourcing Platform

| Evaluator | Criteria | PC | SPC | SPC Iter 1 | FED | Faithfulness |
|---|---|---|---|---|---|---|
| LLM-Eval (Lin and Chen, 2023) | Content | 81.96 | **88.84** | **88.71** | 87.61 | 88.67 |
| | Grammar | 87.12 | **93.64** | **93.68** | 93.09 | 93.56 |
| | Relevance | 86.82 | **94.16** | **93.81** | 92.88 | 93.79 |
| | Appropriateness | 86.99 | **95.84** | 96.17 | 95.68 | **96.19** |
| GPT-Score (Fu et al., 2023) | Fluency | 67.04 | **98.89** | 96.28 | 96.65 | **97.83** |
| | Consistent | 3.47 | **64.25** | **50.43** | 43.45 | 48.69 |
| | Coherent | 69.41 | **100** | **100** | 98.99 | **100** |
| | Depth | 5.40 | **37.36** | **29.30** | 19.40 | 29.01 |
| | Diversity | 72.98 | **96.42** | 94.02 | 92.79 | **94.11** |
| | Likeable | 36.53 | **91.04** | **93.11** | 91.90 | 87.98 |
| G-Eval (Liu et al., 2023) | Relevance (1-5) | 2.288 | **2.992** | 2.986 | 2.941 | **2.99** |
| | Fluency (1-3) | 1.928 | **2.002** | **2** | 1.998 | 1.999 |
| | Consistent (1-5) | 1.736 | **2.651** | **2.587** | 2.449 | 2.496 |
| | Coherent (1-5) | 2.505 | **2.997** | 2.997 | 2.991 | **2.998** |
| | Faithfulness (1-5) | 1.754 | **2.959** | **2.8801** | 2.79 | 2.868 |

Table 16: Results of Automatic Evaluations of Synthetic-Persona-Chat and Persona-Chat. The "FED" column is the evaluation of the dataset generated without FED expert and the column "Faithfulness" is the evaluation results of the dataset generated without the faithfulness expert in the Critic.


Here is a conversation between ***User 1*** and ***User 2***. Please read the conversation and choose all ***self statements*** which describe ***User 2***.

All ***self statements*** mush be inferred from the conversation.

**Conversation:**

User 1: Hi there!
User 2: Hey! How are you?
User 1: I'm doing well. My mom passed away when I was 18. She was from Russia and taught me how to cook some great dishes.
User 2: I'm so sorry for your loss. That's a tough age to lose a parent. What was your favorite dish she taught you?
User 1: My favorite dish she taught me was borscht. It's a really hearty soup that's perfect for the cold winter months.
User 2: That sounds delicious! I've never had it, but I'm definitely going to have to try it now.
User 1: You should! It's really easy to make and it's so filling.
User 2: I'm looking forward to trying it. Thanks for the recommendation!
User 1: No problem! I'm always happy to talk about food.
User 2: Me too! What kind of food do you like to cook?
User 1: I like to cook a variety of foods, but my favorite is probably Italian food.
User 2: Italian food is my favorite too! I love pasta and pizza.
User 1: Me too! I'm a big fan of pasta with red sauce.
User 2: Me too! I could eat pasta every day.
User 1: I could too! It's so good.
User 2: It is! I'm glad we have something in common.
User 1: Me too! We should cook dinner together sometime.
User 2: That would be fun! I'd love to try some of your Russian recipes.
User 1: Great! I'll get started on making a list of recipes.
User 2: Sounds good! I can't wait to try them.
User 1: Me neither

☐ Check it if you think this conversation is toxic.

*Please select all self statements that can describe User 2, based on inferences from the conversation.*

**Select appropriate categories**

| | |
|---|---|
| user2: I like spicy food. | 1 |
| user2: I worked at a movie theater for 4 years. | 2 |
| user2: I m saving up to buy a new camera. | 3 |
| user2: I have never had long hair. | 4 |
| user2: I have always had long hair. | 5 |
| user2: I enjoy running at night. | 6 |
| user2: I m saving up to buy a new car. | 7 |
| None of Them | 8 |

Figure 8: Preview of the Faithfulness Task on the Crowdsourcing Platform.

| Conversation Source | % Lose | $\kappa$ | # annotators |
|---|---|---|---|
| SPC Iter 1 | 17.2 | 0.41 | 50 |
| SPC Iter 2 | 18.5 | 0.48 | 40 |
| SPC Iter 3 | 8.8 | 0.22 | 11 |
| SPC Iter 3* | 8.04 | 0.56 | 24 |
| SPC (LLM2) | 11.5 | 0.49 | 36 |

Table 17: AI detection test results on a sample of 200 conversations. The first column shows the percentage of SPC losing compared to PC in the Turing test. Note that the last iteration (3) of SPC is an evaluation of the segment of conversations based on the extended persona set.

### C.1 Persona Expansion

We assess the importance of the query-based persona expansion module introduced in Section 3.1.1. Similarly to the experiment outlined in Section 4.1, we run the persona expansion on two datasets: Wikipedia and PC. The results of this experiment are presented in Table 18. We designate the persona expansions without the inducted query set (Q) as 'Wikipedia-0', and 'PC-0', and run the same number of iterations for each (100 iterations). We observe that PC-0 includes 4,477 new persona attributes, 20 percent less than PC. The difference in the number of newly generated persona attributes is more pronounced in the case of Wikipedia, where Wikipedia-0 consists of 4,742 persona attributes, 50 percent less than Wikipedia+. This trend is also observed in the number of persona clusters, with PC-0 and Wikipedia-0 having 6% and 49% less clusters respectively. This pattern suggests the effectiveness of the query-based persona expansion in maintaining the diversity of the persona set. Furthermore, the average persona attribute length in PC-0 is 11.38 tokens, which is 28% less than SPC. This reduction points to less detailed and specific persona attributes. In contrast, the expansion in 'Wikipedia-0' exhibits similar average persona attribute lengths compared to 'Wikipedia+'.

### C.2 Conversation Quality

We analyze the effect of the experts within our Critic. We remove each expert, and generate a dataset using one iteration of our framework. We compare the resulting datasets against the output of the first iteration of SPC. We use the evaluators introduced in B.1. The results of this experiment are summarized in Table 16. We observe that the exclusion of the experts results in worse performance according to most criteria: 3 out of 4 in LLM-Eval, 4 out of 6 in GPT-Score, and 3 out of 5 in G-Eval.

### C.3 Faithfulness

We ablate the faithfulness critic, and generate a dataset that we compare against SPC. We compare these datasets both automatically, using human annotators (AI detection test), and using a prompted LLM (LLM-Evaluator). We describe this study in more details below.

**AI Detection Test** We run a human study to compare a small subset of conversations created without the faithfulness expert against their equivalent created with that expert. This experiment process is similar to 4.3 and it is conducted for 200 conversations. The precision decreases from 78.0% to 66.0% without this critic, highlighting its effectiveness in eliminating conversations with contradictory information about user personas. The recall decreases from 36.0% to 23.0%, demonstrating a higher reflection of personas in the conversations in the presence of the faithfulness expert.

**LLM-Evaluator** We extend our comparison to the entire dataset using an LLM as an annotator, following (He et al., 2023; Bansal and Sharma, 2023; Chiang and yi Lee, 2023). Table 19 shows the faithfulness of the conversations generated in the first iteration without the faithfulness expert. The templates used in the LLM-based annotators are described in Table 16 in the rows with "LLM-Faithfulness" as their evaluator. Note that the annotator-based LLM is created using a different LLM, gpt-3.5-turbo (Brown et al., 2020b; Ouyang et al., 2022), than the LLM used for dataset generation.

### C.4 Next Utterance Prediction

We follow the experimental setting described in section 4.2, and compare the performance of various next utterance prediction models trained on SPC against the same models trained on datasets created in the absence of certain experts.

When using the IR Baseline as the next utterance prediction method, we observee that its highest performance of 39% hit@1 occurs when the FED critic is absent during dataset creation. This outcome aligns with FED's emphasis on conversation quality, excluding persona-related aspects. Conversely, the Transformer Ranker, capable of understanding intricate concepts, achieves its peak performance of 13.9% hit@1 when none of the experts are absent. This result supports the inclusion of both FED and the Faithfulness expert in the

| Dataset | PC | SPC | PC-0 | Wikipedia | Wikipedia+ | Wikipedia-0 |
|---|---|---|---|---|---|---|
| # Persona Attributes | 4,723 | **10,371** | 9,200 | 8,768 | **18,293** | 13,510 |
| # Clusters | 323 | **553** | 520 | 408 | **986** | 502 |
| InterCluster-Dist | 0.836 | 0.863 | 0.842 | 0.816 | 0.85 | 0.83 |
| AVG length | 7.65 | **15.9*** | 11.38* | 10.45 | **15.2*** | **15.2*** |

Table 18: Evaluation of the Expanded Persona Attribute Sets. The numbers with '*' indicate the metric value on the newly generated persona attributes, in contrast to the initial persona attributes.

| | LLM Evaluator (%) | | Human Evaluator (%) | |
|---|---|---|---|---|
| Absent Component | Inference | Contradiction | Precision | Recall |
| None | **33.2** | **24.5** | **78.5** | **36.4** |
| Faithfulness | 32.7 | 28.8 | 66.1 | 23.1 |
| FED | 31.7 | 28.5 | N/A | N/A |

Table 19: Faithfulness of Generated Conversation Datasets Using the Framework While Eliminating Each Component. The first row represents the framework without removing any component, equivalent to the first iteration of Synthetic-Persona-Chat.

model architecture. In generative models, the absence of FED impacts the next utterance prediction model the most, leading to a notable decline in performance (e.g. $-12\%$ hit@1, $-9\%$ BLEU, $-10\%$ ROUGE). This observation underscores the crucial role played by FED in enhancing the generative capabilities of the model.

| Absent Component | | Faithfulness | | | FED | | | None | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | Metric | None | Persona | % Change | None | Persona | % Change | None | Persona | % Change |
| IR Baseline | hit@1 | 18.7 | 38.7 | **+106** | 19.0 | **39.0** | +105 | 18.9 | 38.7 | +105 |
| Transformer (Ranker) | hit@1 | 10.9 | 13.5 | +24 | 10.7 | 13.6 | **+27** | 12.4 | **13.9** | +11 |
| Transformer (Generator) | hit@1 | **8.9** | 7.4 | -16 | 8.4 | 7.4 | **-12** | 8.2 | 7.0 | -14 |
| | Perplexity | 204 | 214 | +5 | **174** | 185 | +6 | 203 | 210 | **+3** |
| | BLUE | 0.11 | 0.10 | -11 | 0.11 | 0.10 | **-9** | 0.10 | 0.08 | -15 |
| | ROUGE | 0.14 | 0.15 | -12 | 0.14 | 0.12 | **-10** | 0.13 | 0.10 | -17 |

Table 20: Results of the Next Utterance Prediction Experiment in the Ablation Study. The numbers in the table represent the performance of the trained model on the test portion of the Persona-Chat dataset.