Like hiking? You probably *enjoy nature*: Persona-grounded Dialog with Commonsense Expansions

Bodhisattwa Prasad Majumder[♣] **Harsh Jhamtani**[♠]
Taylor Berg-kirkpatrick[♣] **Julian McAuley**[♣]
[♣]Department of Computer Science and Engineering, UC San Diego
`{bmajumde, tberg, jmcauley}@eng.ucsd.edu`
[♠]School of Computer Science, Carnegie Mellon University
`jharsh@cs.cmu.edu`

Abstract

Existing persona-grounded dialog models often fail to capture simple implications of given persona descriptions, something which humans are able to do seamlessly. For example, state-of-the-art models cannot infer that interest in hiking might imply love for nature or longing for a break. In this paper, we propose to expand available persona sentences using existing commonsense knowledge bases and paraphrasing resources to imbue dialog models with access to an expanded and richer set of persona descriptions. Additionally, we introduce *fine-grained* grounding on personas by encouraging the model to make a discrete choice among persona sentences while synthesizing a dialog response. Since such a choice is not observed in the data, we model it using a discrete latent random variable and use variational learning to sample from hundreds of persona expansions. Our model outperforms competitive baselines on the PERSONA-CHAT dataset in terms of dialog quality and diversity while achieving persona-consistent and controllable dialog generation.

1 Introduction

Persona-grounded dialog generation is a 'chit-chat' dialog setup where a dialog agent is expected to communicate based on a given profile (Zhang et al., 2018). Many recent works have focused on a popular benchmark dataset for this task: PERSONA-CHAT (Zhang et al., 2018) that provides personas as a set of sentences along with each dialog (example in Figure 1). However, a careful analysis of state-of-the-art (SOTA) models reveals that they often struggle to respond to contexts that do not closely match given persona sentences, even when the implications might be obvious to a human.

For example, in Figure 1, the user asks an *indirect* question to the bot related to one of its persona sentences: *I am an animal activist*. SOTA1, which concatenates all persona sentences with dialog history and finetunes a pre-trained generative model (e.g. GPT2) (Wolf et al., 2019), fails to infer implied commonsense from the dialog context and conditions on an incorrect persona. SOTA2, which separately selects a persona sentence given the dialog history (Lian et al., 2019) manages to choose the correct persona but merely copies it as the final response. Neither approach is in general capable of responding to context that goes beyond what is explicitly mentioned in the available persona sentences, which limits consistent and interesting conversation. The goal of our model is to understand that being 'an animal activist' may imply that the person wants 'to make a difference' via their activity towards animals and synthesizes a context-consistent and engaging response.

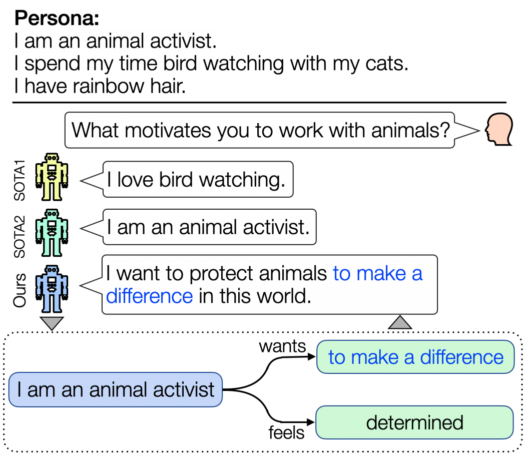

Figure 1: State-of-the-art models struggle to respond a user's query, where generating an engaging response depends on commonsense reasoning.

In this paper, we focus on making persona-grounded chatbots more consistent with personas and implicit dialog context. We present a framework to expand available persona sentences to their commonsense implications by using an existing

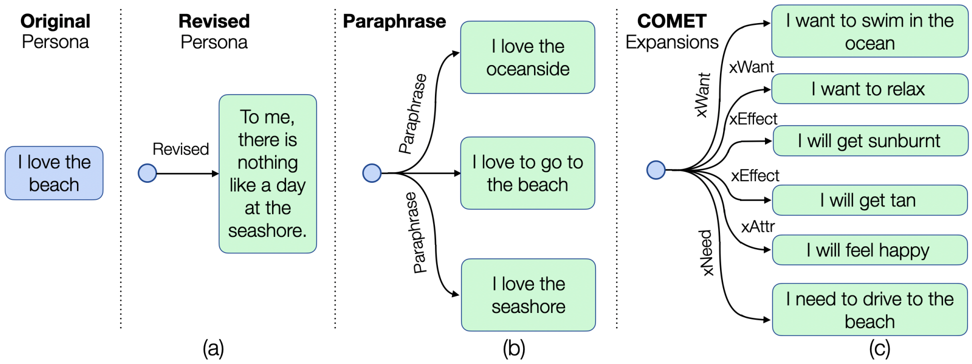

Figure 2: Expansions of an original persona via (a) human rewrite (2018), (b) paraphrase, and (c) COMET.

commonsense knowledge base or paraphrasing resources (see Section 3). We endow our dialog model with these expansions directly rather than requiring the model to learn them from scratch for being context-consistent. We find that expansions derived from a commonsense knowledge base are more useful to provide engaging contextual information compared to other expansion sources.

We further propose a **Com**mon Sense and **P**ersona **A**ligned **C**hatbot[1] (**COMPAC**) which models choices over the *expanded* persona set via a discrete latent random variable (See Section 4) as *fine-grained* persona grounding. Even though it is tractable to marginalize over all expansions, that would require a forward pass through the dialog generator for each outcome which is prohibitively slow during training. Instead, to accommodate hundreds of persona expansions, we train the model by optimizing a lower bound on the log-likelihood. We use amortized variational inference by approximating the true posterior using an inference network that eventually provides useful inductive bias. Particularly, we show that our Bayesian formulation for the fine-grained persona grounding was essential as simply providing expanded knowledge does not help the model generate better responses.

We also outperform competitive baselines in all dialog quality metrics as well as human evaluations which find COMPAC to be engaging and coherent. We demonstrate that COMPAC learns to be consistent with the dialog context with accurate persona grounding especially in the presence of commonsense expansions. Finally, we show that our model can reflect a change in response generation when a grounding persona is modified, indicating the possibility of controllable generation.

[1]Code is available at – https://github.com/majumderb/compac.

2 Persona Grounded Dialog

We use a popular benchmark dataset: PERSONA-CHAT (Zhang et al., 2018) for our persona-grounded dialog generation task. It contains 10,907 dialogs between pairs of speakers where each speaker follows their own persona; 968 dialogs are used for validation and 1,000 for testing. Each speaker is described by 3-5 persona sentences. (e.g. 'I love the beach', 'My mother is a medical doctor'). Out of 1,155 total unique personas, 100 are used for validation and 100 for testing.

The task of persona-grounded dialog generation is: given a dialog history H and grounding persona sentences S, we must predict the next utterance x (Summary of notations in Table 1). Hence a dialog model should maximize the likelihood $p(x|H, S)$. From the PERSONA-CHAT dataset, we use 131,438 utterances for training the dialog model, 15,602 for validation, and 15,024 for testing.

3 Persona Expansion

Persona sentences used in persona-grounded dialogs are instances of world events that often imply real-world consequences or richer information. For example, 'I love surfing' naturally implies that the person might be 'adventurous' or 'loves the outdoors'. Similarly, it also means that the person wants 'to go to the beach' regularly. Inferring these *expansions* from the original fact is non-trivial without additional commonsense knowledge.

Zhang et al. (2018) found evidence that having human written interpretations of a persona sentence via rephrasing often helps in providing novel information in persona grounding. While obtaining such expansions by manual rewriting is expensive, here we explore two automatic ways to generate them at scale and separately evaluate them on the downstream dialog modeling task.

3.1 COMET

COMET (Bosselut et al., 2019) is a framework that generates rich and diverse commonsense expansions of a given world event. It is a finetuned version of a pre-trained GPT2 (Radford, 2018) model on a pre-existing commonsense knowledge graph such as ATOMIC (Sap et al., 2019) that can generate novel nodes (events) and edges (relations), as seen in Figure 2c. Specifically, ATOMIC provides tuples that belong to nine relation types spanning over cause-effect interrelations between events: `oEffect`, `oReact`, `oWant`, `xAttr`, `xEffect`, `xIntent`, `xNeed`, `xReact`, and `xWant`—where a prefix '`x`' indicates an effect or cause on the person and '`o`' denotes the same on others. While we tried COMET finetuned on an alternative commonsense knowledge base (e.g.) ConceptNet, not all of the expansions were appropriate to describe a persona, mainly because we observe that persona sentences are *event*-like ('I love to go to the beach') as opposed to *concepts* such as 'beach'. For more details on COMET and ATOMIC we refer the reader to (Bosselut et al., 2019) and (Sap et al., 2019) respectively.

We use the COMET framework to generate expansions for each persona sentence along the nine relation types that ATOMIC provides. We obtain different samples while decoding via beam search from COMET for more diverse and unique expansions, as shown in Figure 2c. We preprocess these expansions to add suitable prefixes to make them similar to the original persona. For example, expansions relating to `xWant` and `xAttr` are prefixed with 'I want' and 'I am' respectively. For each persona sentence, we generate 5 expansions per relation, i.e., in total we will obtain $5 \times 9 = 45$ expansions per persona sentence.

3.2 Paraphrasing

To explore alternative sources for generating commonsense expansions beyond COMET, we consider paraphrasing persona sentences. Paraphrases of a sentence convey almost the same meaning to a listener as the original. Often paraphrases use synonymous phrases or manipulate word-syntax of the original sentence, which implicitly involves both context comprehension and world knowledge (Zeng et al., 2019). We obtain these in two ways:

Paraphrase Network To generate paraphrases at scale, we use an off-the-shelf paraphrasing system based on back-translation (Xie et al., 2019; Federmann et al., 2019) with pre-trained language translation models. We make use of `En-Fr` and `Fr-En` pre-trained translation models as the components for back-translation.[2] While we tried other language pairs, the `En-Fr` pair proved the most satisfactory based on qualitative analysis on 500 samples. We generate 5 paraphrases per persona sentence, which readily provides more lexical and syntactic variants as shown in Figure 2b.

Manual Paraphrasing To compare with other expansions, we reuse manually written revised versions of persona sentences provided with PERSONA-CHAT (Zhang et al., 2018) though these are limited to only one paraphrase per sentence. We call them **revised** for short (see Figure 2a).

4 Common sense and Persona Aligned Chatbot (COMPAC)

To infuse commonsense context in persona-grounded dialog generation, we imbue our dialog model with the expanded persona set instead of only original personas S. But these persona expansions lead to hundreds of new sentences as opposed to only a few given persona sentences which makes it infeasible to encode using a single transformer, as was done in prior works (Wolf et al., 2019). Additionally, encoding all persona sentences as a single text input leads to a lack of interpretability i.e., it is not clear which persona sentence was used by the model in generating a particular response.

Instead, we propose COMPAC: **Com**mon Sense and **P**ersona **A**ligned **C**hatbot that allows us to make a *fine-grained* choice of a persona sentence to generate the target response. Let C denote a list of expended personas, derived from S (including S itself). We further add a null persona $\varnothing$ in C considering that some utterances can purely condition on the dialog context. We are interested in modeling the conditional $p(x|H,C) = p(z|H,C)p(x|z,H,C)$ where $z \in \{1, 2, \ldots, |C|\}$ is a latent discrete random variable, unobserved in the data. Given the dialog history H, first we sample a particular persona sentence C_z from a *prior network* $p_\theta(z|H)$ (see Figure 3). Next, as depicted in Figure 3, the dialog response x is sampled from a *generator network* $p_\phi(x|H, C_z)$ by conditioning on the history H and chosen persona sentence C_z.

In the generative model described above, the latent variable z is a discrete random variable which

[2] https://github.com/google-research/uda

S	Set of original persona sentences
C	Set of expanded persona sentences (includes S and a null persona $\varnothing$)
H	Dialog history with alternative turns from each speaker
x	Target utterance
z	Discrete latent random variable $\in \{1, 2, \ldots, \|C\|\}$
e	Mean of RoBERTa subword embeddings as an encoder
t_k	Expansion type for k-th expansion
f_i	i-th feature function for prior network; $i \in \{1, 2, 3\}$
θ	Parameters for prior network $p_\theta(z\|H, C)$
ϕ	Parameters for generator network $p_\phi(x\|H, C_z)$
α	Parameters for inference network $p_\alpha(z\|x, H, C)$

Table 1: Summary of notation used in the paper

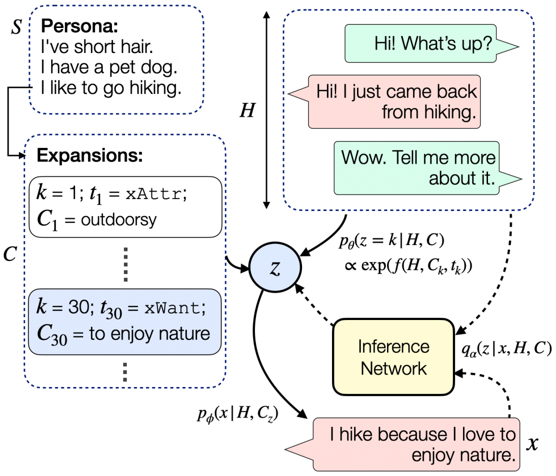

Figure 3: COMPAC samples a persona sentence from the prior and generates the response conditioned on the dialog context and sampled persona. The inference network is used only during training.

points to a single persona sentence. This decision (of conditioning on a single persona sentence) was based on the observation that most dialog responses in the datasets under consideration are relevant to only one persona sentence. It is possible to allow for multiple persona sentences by defining z to pick a subset of $|C|$ persona sentences instead of picking a single sentence. We leave this as a possible future extension.

4.1 Persona Choice Prior

The dialog history H can hold cues regarding which persona sentence might be applicable given the context. For example, in Figure 3 the historical context suggests that 'following fashion trends' can be a consequence of 'being fashionable'.

We encode both the dialog history H and persona sentence C_k by averaging RoBERTa subword embeddings (Liu et al., 2019) as $e(H)$ and $e(C_k)$. We use an implementation from HuggingFace for RoBERTa[3] with `roberta-base` as the pretrained model. Then we parameterize the prior $p_\theta(z|H, C)$ as a log-linear model with the following features:

Dialog history We obtain $f_1(H, C_k)$: a scalar feature using a bilinear product $\langle e(H), e(C_k)\rangle$ to align the persona sentences with the dialog history.

Expansion types Each k-th persona expansion corresponds to an expansion type t_k. In the case of COMET, these types are the nine commonsense relations provided by ATOMIC (see Section 3.1). For paraphrased expansions, we annotate each as type `paraphrase` and the original persona sentences as `original`. We consider two additional features with expansion types: (a) $f_2(t_k)$ that represents a global preference over the relation type embedded via a type embedding layer; and (b) $f_3(t_k, H)$ that appends the expansion type embedding with dialog history encoding $e(H)$, followed by a linear layer to obtain a real-valued score for history-specific preference over the expansion type.

The dimension of the expansion type embedding was set to 5. Finally, the prior model can be represented concisely as $p_\theta(z = k|H, C) \propto \exp(f(H, C_k, t_k))$, where $f(H, C_k, t_k)$ is the sum $\lambda_1 * f_1(H, C_k) + \lambda_2 * f_2(t_k) + \lambda_3 * f_3(t_k, H)$ with λ_i's are trainable parameters.

4.2 Generator Network

Following prior work (Wolf et al., 2019), we use pre-trained GPT2 (Radford, 2018) (Transformer with 12 layers, 768 hidden size, 12 heads— `gpt2-small`[4]) to generate dialog responses given the dialog history H, with the selected persona sentence C_z prepended to it. In the case of C_z being the null persona, an empty string is prepended. We further append the target response x to the combined context $(C_z; H)$, and feed the sequence to GPT2, after tokenization. To distinguish between persona tokens, history tokens, and target response tokens, we use segment indicators—{`Persona`, `Speaker1`, `Speaker2`}—for which corresponding embeddings are learned via a separate segment embedding layer in the model. We add the segment embedding to the corresponding token embedding in the model input layer. To obtain the conditional likelihood $p_\phi(x|H, C_z)$, we only consider the target tokens

[3] https://huggingface.co/transformers/model_doc/roberta.html

[4] https://github.com/huggingface/transfer-learning-conv-ai

for cross-entropy loss calculation.

Wolf et al. (2019) also leveraged incorrect responses given a dialog history from PERSONA-CHAT as negative samples in an auxiliary loss to encourage the correct candidate to obtain the highest likelihood compared to the incorrect ones. However, we did not find any improvement using this loss in COMPAC.

4.3 Learning and Inference

Our training data $\mathcal{D}$ consists of instances of dialog history H and ground truth dialog responses x. We train our model parameters θ and ϕ to maximize the likelihood of the target dialog response x given the dialog history: $\log p(x|H, C; \theta, \phi)$ totalled over $\mathcal{D}$. Since the discrete random variable z is unobserved in the training data, we must marginalize over z to compute the desired likelihood $p(x|H; \theta, \phi)$:

$$\log p(x|H; \theta, \phi) = \log \mathbb{E}_{z \sim p_\theta(z|H)}[p_\phi(x|z, H)];$$

where we drop C from the conditionals for simplicity.

Inference Network Note that the number of persona expansions is typically in the range 150-250, and thus it is computationally expensive to marginalize over the entire selection space of z during training. We instead optimize a variational lower bound (ELBO) of $\log p(x|H; \theta, \phi)$ given as

$$\mathbb{E}_{z \sim q_\alpha(z|H)}[\log p_\phi(x|z, H)] \\ - KL(q_\alpha(z|x, H)||p_\theta(z|H)),$$

where we use the inference network $q_\alpha(z|x, H)$ to compute the approximate posterior (Kingma and Welling, 2014). In our initial experiments, we observe that using an inference network leads to better perplexity values than using samples from the prior.

The architecture of the inference network is similar to that of the prior network, a log-linear model. Along with the features related to dialog history and expansion types, we additionally include another scalar feature: a bilinear product $\langle x, C_k \rangle$ between the encoded persona and ground truth response x encoded with RoBERTa embeddings to align the persona choice according to the target utterance.

Optimization The parameters of the generator network (ϕ) and prior network (θ) are trained directly via back-propagation. Since z is a discrete latent variable, we use REINFORCE (Williams, 1992) to train the inference network parameters α. However, the REINFORCE estimator often suffers from high variance. To reduce the variance, we found it useful to (1) use a moving average baseline (Zhao et al., 2011); and (2) regularize the prior network by penalizing the entropy of the output categorical distribution. To avoid KL mode collapse, we use KL-annealing (Bowman et al., 2016) where we linearly increase the weight of the KL term beginning from 0 to 1 as training progresses.

Decoding At decoding time, we first sample k from the prior $p_\theta(z|H, C)$, and then C_k is fed to the generator network. Following previous work (Wolf et al., 2019), we use nucleus sampling (Holtzman et al., 2020) (with $p = 0.95$) to decode the final response from the probabilities produced by the generator. We also found that high-temperature sampling from the prior often leads to more diverse generation.

5 Experiments

We conduct our experiments based on the following desiderata: (1) Do persona expansions help to generate high quality and diverse responses? (2) Does COMPAC achieve accurate persona grounding given a dialog context? (3) Does COMPAC enable persona-consistent and controllable generation? Hyperparameter details are in Appendix §A.

5.1 Baselines

To demonstrate the efficacy of COMPAC, we compare it with three competitive baselines on the PERSONA-CHAT dataset:

1. **Per-CVAE:** A CVAE model that exploits persona sentences for diverse generation with an external memory (Song et al., 2019b)

2. **LIC + KS:** The best performing transformer model (Lost in Conversation i.e., LIC) in terms of human evaluation in the ConvAI2 NeurIPS competition (Dinan et al., 2019a) combined with a knowledge-selection (KS) mechanism Lian et al. (2019) to achieve state-of-the-art results on PERSONA-CHAT;

3. **GPT2:** Finetuned GPT2 on PERSONA-CHAT just by concatenating all persona sentences along with dialog history (Wolf et al.,

System	PPL	BLEU-1	BLEU-2	D-1	D-2
Original					
Per-CVAE (2019b)	48.37	0.19	0.11	0.03	0.21
LIC + KS (2019)	30.50	0.18	0.07	0.07	0.24
GPT2 (2019)	21.46	1.42	0.78	0.05	0.11
COMPAC-original	19.56	3.24	1.31	0.15	0.25
Paraphrased					
GPT2-revised	21.01	1.54	0.97	0.13	0.25
GPT2-paraphrase	21.57	1.61	0.86	0.16	0.35
COMPAC-revised	18.12	3.52	0.99	0.48	0.65
COMPAC-paraphrase	17.09	3.83	**1.87**	0.56	0.85
COMET					
GPT2-COMET	21.12	1.62	0.81	0.21	0.39
COMPAC	**16.21**	**4.12**	1.82	**0.87**	**1.07**

Table 2: Dialog quality metrics on the PERSONA-CHAT test set. PPL=Perplexity, D-1/2=% of distinct uni- and bi-grams.

Persona: I enjoy listening to classical music. I'm a Hindu. My favorite color is red.
User: Hi, recently I have got interests in religion.
GPT2 (2019)**:** Hi! How are you?
COMPAC-original: I'm a Hindu.
COMPAC-revised: Hi! I am a Hindu too.
COMPAC-paraphrase: That's great. I am religious.
COMPAC: That's great. I go to temple regularly and learn about Hinduism.

Table 3: Sample generations by different models. More examples are in Appendix §C.

COMPAC vs.	GPT2 (2019)		LIC + KS (2019)		GPT2-COMET		COMPAC-og		COMPAC-par		Gold	
Metric ↓	win	loss	win	loss	win	loss	win	loss	win	loss	win	loss
Fluency	**81.2***	5.1	**83.2***	6.7	**90.5***	2.3	**68.0**	26.0	**65.0**	19.4	40.1	**42.5**
Engagement	**90.5***	3.3	**87.4**	5.9	**97.6***	0.5	**86.5***	10.5	**81.5***	10.5	**62.1***	30.5
Relevance	**78.2***	4.8	**78.0***	7.7	**93.2***	1.8	**65.5***	18.5	**62.1**	15.6	32.8	**54.6***

Table 4: Pairwise comparison between responses generated by COMPAC vs. responses generated by other baselines (og: original, par: paraphrase) as well as the Gold response. All numbers are in percentages with **bold** indicates the highest. Ties are not shown. Entries with * denote significance with $p < 0.05$ from bootstrap tests on 1000 subsets of size 50.

2019) to obtain the best automatic metric in the ConvAI2 competition.

A minimal version of COMPAC is also considered, **COMPAC-original**, which only uses the original persona, for a direct comparison with other model architectures that only use the original persona. Furthermore, to justify the choice of fine-grained persona grounding for an effective utilization of persona expansions, we also consider baseline versions of GPT2 trained with each of the expansion strategies: **GPT2-revised**, **GPT2-paraphase**, and **GPT2-COMET**. To show that COMPAC can work with persona expansions derived from various sources, we compare with versions of COMPAC trained with paraphrase-based expansions: **COMPAC-revised** and **COMPAC-paraphrase**. By default, COMPAC indicates it is trained with COMET expansions.

5.2 Comparison of Dialog Quality

We measure perplexity for language modeling performance, and BLEU-1 (Papineni et al., 2002) and BLEU-2 (Vedantam et al., 2015) scores between generated and gold utterances to measure the fidelity of the generated responses. Given our goal of generating engaging responses with novel information, we deem it important to consider the diversity in the generated responses which we measure using D-1 and D-2 (percentage of distinct uni- and bi-grams respectively) (Li et al., 2016).

Table 2 shows that COMPAC outperforms three competitive baselines when trained on the original persona in all quality metrics indicating the efficacy of our architecture. Moreover, when combined with persona expansions, we observe a modest 3-8 point decrease in perplexity and a large improvement in both BLEU and diversity scores which confirms that COMPAC successfully leverages the persona expansions to improve dialog quality. COMPAC trained with COMET expansions achieves the best performance both in terms of fidelity and diversity which shows that COMET expansions help the model to respond to implicit context with commonsense and to explore novel information. But with revised personas, we find that both COMPAC and GPT2 provide marginal performance gains, mirroring the observation from (Zhang et al., 2018). Finally we observe gradual degradation in performance when we trivially finetune GPT2 with paraphrase and COMET expansions. Note that GPT-2 could have implicitly learned to focus on a single persona attribute. However, the COMPAC performs better suggesting that fine-grained persona grounding acts as a useful inductive bias in effectively utilizing larger expansion sets.

5.3 Human Evaluation for Dialog Generation

Automatic evaluation of dialog systems is still notoriously unreliable (Liu et al., 2016; Novikova et al., 2017) and such systems should be evaluated by human users. Hence, we perform pairwise comparisons between responses generated our best system, COMPAC trained on COMET expansions, and responses generated by four strong baselines: GPT2, GPT2-COMET, COMPAC-original, COMPAC-paraphrase (the best COMPAC model with paraphrase expansions). We also consider the gold responses for comparison. We conduct a human evaluation with 100 test examples on three aspects critical for practical use: (1) **Fluency** measures whether the generated output is fluent (in English); (2) **Engagement** measures whether the generated response is engaging or interesting; and (3) **Relevance** measures whether the generated output is relevant with respect to the dialog history. More details of the evaluation are in Appendix §B.

Table 4 shows that human annotators found responses generated by COMPAC trained with COMET expansions more engaging as compared to responses from all the baselines as well as the gold responses by statistically significant margins. This confirms our hypothesis that COMET expansions were helpful in adding novel content. Human judges also found that despite a significant drop in perplexity, COMPAC was not more fluent than COMPAC-original and COMPAC-paraphrase with statistical significance, indicating similar language modeling performance. We find the inter-annotator agreement, as measured by Cohen's kappa (Cohen, 1960), for fluency, engagement, and relevance were 0.62, 0.71, and 0.73 respectively.

5.4 Fine-grained Persona Grounding

Next we want to investigate the extent of COMPAC's ability to ground the response generation with a fine-grained persona choice as a probing experiment. Specifically, we want to measure whether our model can choose a coherent persona from the available persona sentences given the dialog context. Note that in persona-grounded chitchat, not all utterances are tied to a personas and could be purely based on dialog context. We find that 44% of the time the model selects the null persona (∅) and conditions only on the dialog history. To assess the persona grounding for the remaining (56%) utterances, we perform (a) a persona entailment experiment, and (b) human evaluation.

System	Persona Entailment Prior	Persona Entailment Inference Network	Human eval.
Original			
COMPAC-original	25.5	79.3	–
Paraphrased			
COMPAC-revised	20.6	78.9	40.6
COMPAC-paraphrase	27.8	87.3	67.8
COMET			
COMPAC	**37.9**	**96.4**	**87.3**

Table 5: Assessment of persona grounding with and without inference network using the DNLI entailment set. Human evaluation (eval.) was conducted to measure the relevance when an expanded persona is chosen–all entries are statistically significant.

Persona Entailment We adapt the Dialogue Natural Language Inference (DNLI) dataset (Welleck et al., 2019) and collect persona-utterance pairs that belong to an *entailment* relation. This results in a subset of 4,613 utterances with associated ground truth persona sentences in our test set. Next, we obtain a persona sentence by performing argmax over the prior $p_\theta(z|H, C)$ as well as the inference network $q_\alpha(z|x, H, C)$ from our COMPAC models and calculate accuracy with the ground truth persona. For models that use expanded personas, we track the original persona from the retrieved expansion for accuracy calculation. Table 5 shows that COMPAC with COMET achieves the most accurate persona grounding suggesting that inference networks can approximate the true posterior better when a commonsense persona is available for grounding. In the case of the prior, a better entailment accuracy than random chance $(1/5)$ confirms our choice of the history-conditioned prior network rather than a uniform prior.

Human Evaluation Since DNLI does not entail expanded personas, we conduct a human evaluation to judge the relevance of a chosen persona *expansion* sampled from the inference network. Specifically, we ask: *Is this knowledge relevant to the given dialog history?*—with options as 'Yes', 'No', and 'Uncertain'—and with 100 examples (more in Appendix §B) for each COMPAC variant that uses expanded personas. The inter-annotator agreement, as measured by Cohen's kappa was 0.76.

System	Unigram Overlap Recall	Precision	F1	BERT Score
Original				
LIC + KS (2019)	10.4	34.2	15.3	–
COMPAC-original	14.9	39.1	21.6	57.2
Paraphrased				
COMPAC-revised	15.2	40.3	22.1	58.1
COMPAC-paraphrase	17.8	42.2	25.1	72.9
COMET				
COMPAC	**21.4**	**48.9**	**29.8**	**78.8**

Table 6: Conditional generation performance on the PERSONA-CHAT test set to show the similarity between generated responses and grounding persona sentences. We omit GPT2-based models since they do not select a particular persona sentence for grounding.

Performance	Example
Presence of changed entity **86%**	**Changing the key entity** Before: My favorite color is red After: My favorite color is **green**
	Conversation: User: What is your favorite color? Bot: My favorite color is **green**
BERT score with unedited persona: 46.2 edited persona: **74.6**	**Swapping with another expansion** Before: I want to swim in the ocean After: I want to buy a beach **umbrella**
	Conversation: User: What do you do at beaches? Bot: I will buy an **umbrella** at the beach

Table 7: Controlled generation with edited persona

Again, Table 5 shows that models with COMET expansions can choose the most relevant persona sentence which corroborates our claim in persona entailment experiments. On average, we noticed that COMPAC with COMET expansions prefers to choose expanded personas 87% of the time out of all non-null persona choices. This reduces to 62% in the case COMPAC-paraphrase. In contrast, COMPAC-revised tends to select an original persona over an expansion more often.

5.5 Controllable Generation

Controllable generation of persona-grounded dialog can help to generalize the dialog agent to newer persona details just by changing the grounding in the conditional generator. While controllable text generation with a desired attribute has gained interest recently (Dathathri et al., 2020; Kong et al., 2019), we investigate the possibility of controlling generation with a desired persona and measure the performance of the conditional generator. For this, we observe a set of knowledge overlap metrics—the unigram recall/precision/F1 scores–from Dinan et al. (2019b) and BERT score (Zhang et al., 2020) for semantic similarity between the generated responses and the persona retrieved. Table 6 shows that conditional generation is strongest when COMPAC is trained with COMET suggesting commonsense expansions are more appropriate to the dialog context in influencing the response generation.

Next, we create a diagnostic dataset of 100 examples where we manually edit the persona by changing an entity in a persona sentence or swapping the selected persona expansion with another relevant one (See examples in Table 7) to directly measure controllability in response generation. We observe that COMPAC can successfully reflect the entity-change in the generated response based on the change in the persona grounding 86% of the time. For a swapped persona expansion, we also see a higher BERT score (74.6) between the edited persona and newly generated response as opposed to a lower score (46.2) with the unedited persona. Together with the qualitative examples in Table 7 this suggests that COMPAC supports controllable generation with contextually modified personas.

5.6 Qualitative Analysis

Table 3 shows responses from different models for a sample dialog context. Qualitatively, we find that COMPAC with COMET expansions responds to the context with commonsense using novel content from a commonsense expansion (being Hindu $\rightarrow$ to learn about Hinduism), where other responses remain generic or incoherent. In Table 8, we illustrate responses generated by the COMPAC model along with the underlying persona choice sampled from the prior network. Cases show that COMPAC successfully chooses an original or an expanded persona sentence, as appropriate, but also defaults to the null persona ($\varnothing$) that leads to a bland response.

6 Related Works

Building personalized dialog agents has been a popular task recently, thanks to Zhang et al. (2018) who extensively studied the task with a new dataset PERSONA-CHAT, later as a form of a challenge (Dinan et al., 2019a), where the dialog agent is seeded with a predefined persona in the form of multiple sentences of textual description, mirroring a casual human conversation which many times draws snippets from individual personal experi-

Verdict	Persona	Dialog History	Persona Choice (z)	Response
Uses Original Persona	1. I'm the youngest of five. 2. I work at the hospital as a nurse. 3. I've pink hair.	What do you do for work?	I work at the hospital as a nurse.	I am a nurse.
Uses Expanded Persona	1. I just want to have fun with my friends. 2. I don't drink or do drugs or anything. 3. I am 19 and I cannot wait to move out of my parents home.	Are you enjoying life?	I just want to have fun with my friends → wants to have a party	Not really. I want to have a party.
Uses Improper Persona	1. I make a million dollars a year. 2. I'm married and have three kids. 3. I'm a baseball player.	I find it hard to support my family working at a bar. What about you?	Null persona ($\varnothing$)	I enjoy my life.

Table 8: Examples showing correct and incorrect persona choices in various dialog contexts by COMPAC model. It shows that COMPAC is capable of choosing a correct persona sentence (original or expanded) but sometimes the prior network fails to sample an appropriate one (third case).

ences and facts. Recent works focus on improving persona-grounded dialog generation performance (Wolf et al., 2019; Mazaré et al., 2018; Bao et al., 2019) as well as persona consistency in generated dialog (Welleck et al., 2019; Li et al., 2019; Song et al., 2019a). Bao et al. (2019) proposed a reinforcement-learning-based framework that promoting informativeness and persona-consistency via personal knowledge exchange. Xu et al. (2020) focused on using plausible topical keywords related to the available persona facts using a neural topic model to explore beyond the given knowledge, possibly closest to our work. We rather focus on obtaining commonsense implications of the given persona in the form of text snippets that are more expressive than topical keywords.

Persona-grounded dialog generation is a special case of knowledge-grounded dialog generation. Knowledge grounding in dialog has many real-world applications that are well-studied in recent literature (Zhou et al., 2018; Ghazvininejad et al., 2018; Dinan et al., 2019b; Lewis et al., 2019). In this work we use fine-grained grounding/selection on persona which performed better than encoding the entire persona for each response. Such fine-grained selection has been found useful in prior works on text generation such as dialog (Lian et al., 2019) and image captioning (Jhamtani and Berg-Kirkpatrick, 2018). For dialog generation, a contextual knowledge selection has been successfully applied in prior works (Parthasarathi and Pineau, 2018). Specifically, Zhao et al. (2017) and later Song et al. (2019b) proposed a conditional-VAE framework to learn latent context given the dialog history to guide knowledge selection.

Finally, few recent works focused on augmenting grounding with commonsense knowledge with successful applications in open-domain topical dialog generation (Ghazvininejad et al., 2018; Moon et al., 2019), story generation (Mao et al., 2019) and sarcasm generation (Chakrabarty et al., 2020). In this work, we extend this effort into persona-grounded dialog generation via augmenting grounding persona with commonsense knowledge.

7 Conclusion

In this work, we showed that expanding persona sentences with commonsense helps a dialog model to generate high-quality and diverse persona-grounded responses. Moreover, we found that *fine-grained* persona grounding is crucial to effectively condition on a large pool of commonsense persona expansions, which further provided additional controllability in conditional generation.

While our expansions are limited by the performance of COMET or paraphrase systems, we envision future work to train the dialog model end-to-end along with the expansion generation. As future work, we would like extend the prior network to sample more than one persona sentences by expanding the sample space of the discrete random variable to generate more interesting responses.

Acknowledgments

We thank Arthur Szlam, Y-Lan Boureau, Michel Galley, Sujoy Paul, and anonymous reviewers for providing valuable feedback. BPM is partly supported by NSF Award #1750063. HJ is partly supported by a Adobe Research Fellowship. Findings and observations are of the authors, and do not necessarily reflect the views of the funding agencies.

References

Siqi Bao, Huang He, Fan Wang, Rongzhong Lian, and Hua Wu. 2019. Know more about each other: Evolving dialogue strategy via compound assessment. In *ACL*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. In *ACL*.

Samuel Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *SIGNLL*.

Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. 2020. R^3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. *CoRR*, abs/2004.13248.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander H. Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander I. Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2019a. The second conversational intelligence challenge (convai2). *CoRR*, abs/1902.00098.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019b. Wizard of wikipedia: Knowledge-powered conversational agents. In *ICLR*.

Christian Federmann, Oussama Elachqar, and Chris Quirk. 2019. Multilingual whispers: Generating paraphrases with translation. In *W-NUT@EMNLP*.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *AAAI*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *ICLR*.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to describe differences between pairs of similar images. In *EMNLP*.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *ICLR*.

Xiang Kong, Bohan Li, Graham Neubig, Eduard H. Hovy, and Yiming Yang. 2019. An adversarial approach to high-quality, sentiment-controlled neural dialogue generation. *CoRR*, abs/1901.07129.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL HLT*.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2019. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. *CoRR*, abs/1911.03860.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *IJCAI*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Huanru Henry Mao, Bodhisattwa Prasad Majumder, Julian J. McAuley, and Garrison W. Cottrell. 2019. Improving neural story generation by targeted common sense grounding. In *EMNLP*.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *EMNLP*.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *ACL*.

Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *EMNLP*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Prasanna Parthasarathi and Joelle Pineau. 2018. Extending neural generative conversational model using external knowledge sources. In *EMNLP*.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *AAAI*.

Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. 2019a. Generating persona consistent dialogues by exploiting natural language inference. *CoRR*, abs/1911.05889.

Haoyu Song, Weinan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019b. Exploiting persona information for diverse generation of conversational responses. In *IJCAI*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

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

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4).

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, abs/1901.08149.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. 2019. Unsupervised data augmentation. *CoRR*, abs/1904.12848.

Minghong Xu, Piji Li, Haoran Yang, Pengjie Ren, Zhaochun Ren, Zhumin Chen, and Jun Ma. 2020. A neural topical expansion framework for unstructured persona-oriented dialogue generation. *CoRR*, abs/2002.02153.

Daojian Zeng, Haoran Zhang, Lingyun Xiang, Jin Wang, and Guoliang Ji. 2019. User-oriented paraphrase generation with keywords controlled network. *IEEE Access*, 7.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *ACL*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *ICLR*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*.

Tingting Zhao, Hirotaka Hachiya, Gang Niu, and Masashi Sugiyama. 2011. Analysis and improvement of policy gradient estimation. In *NIPS*.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*.

A Implementation Details

We obtain the PERSONA-CHAT dataset from ParlAI repository[5]. For COMET expansions, we use the code[6] released by the authors of COMET (Bosselut et al., 2019). We performed BPE tokenization with the GPT2Tokenizer[7].

Network architectures For the generator network, we use GPT2 (Transformer with 12 layers, 768 hidden size, 12 heads— `gpt2-small`[8]) following the state-of-the-art model (Wolf et al., 2019) from Conv-AI2 competition. Wolf et al. (2019) also leveraged incorrect responses given a dialog history from PERSONA-CHAT as negative samples in an auxiliary loss to encourage the correct candidate to obtain the highest likelihood compared to the incorrect ones. However, we did not find any improvement using this loss in COMPAC. COMPAC has total of 164 Million parameters whereas GPT2 based baseline has 124 Million parameters.

Hyperparameters Following (Wolf et al., 2019) we use history size 2 (i.e. 4 previous utterances). We use AdamW optimizer (Loshchilov and Hutter, 2017) and the learning rate was set at $6.25e-5$ with a linear decay of step size 10^{-1} per epoch. The baseline in REINFORCE was done with a discounted moving average with a ratio of 0.99. The REINFORCE loss coefficient was set at 0.8 and the language modeling loss coefficient was set to 1.0.

Training Each model converged in 3 epochs on an average with batch size 4 in a TITAN X (Pascal) GPU that took 12 hours in total. While training, we only observe perplexity on the validation set to employ an early-stopping criteria.

---

[5] http://parl.ai/downloads/personachat/personachat.tgz

[6] https://github.com/atcbosselut/comet-commonsense

[7] https://huggingface.co/transformers/model_doc/gpt2.html

[8] https://github.com/huggingface/transfer-learning-conv-ai

B Evaluation

Automatic Evaluation During dialog quality evaluation, perplexity is measured by adapting the official evaluation protocol from the Conv-AI2 challenge[9].

To assess persona grounding, we use Dialogue Natural Language Inference (DNLI) dataset (Welleck et al., 2019) that has persona-utterances pairs under three classes—entailment, neutral, and contradiction. We gather all the entailment pairs including all splits that resulted in 44,000 persona-utterance pairs. Then we map with the PERSONA-CHAT test set to obtain 4,613 utterances associated with a ground truth persona.

For assessing conditional generation performance, we use BERT score from the publicly available repository[10].

Human Evaluation For human evaluation, we hired two Anglophone (Lifetime HIT acceptance $\% > 80$) annotators for every human-evaluated test generation. Figure 4 shows a sample question for a human judge for the pairwise comparison of a response generated by COMPAC and a response generated by a baseline for three aspects—fluency, engagement, and coherence.

While measuring persona grounding, we used a similar setup where we provided a dialog history and a sampled expansion and asked 'Is this knowledge relevant to the given dialog history?'— with three options —'Yes', 'No', and 'Uncertain' (See Figure 5). Similar to the previous human evaluation study, we hired two Anglophone (Lifetime HIT acceptance $\% > 80$) annotators for each question. We find the inter-annotator agreement, as measured by Cohen's kappa was 0.76.

C Generation Examples

Tables 9 to 12 present generations from COMPAC for sampled dialog histories with comparison across baselines.

[9] https://github.com/facebookresearch/ParlAI/blob/master/projects/convai2

[10] https://github.com/Tiiiger/bert_score

Persona: I've short hair. I have a pet dog. I like to go hiking.
User: Do you follow current fashion trends?
GPT2 (2019): Yes. What about you?
COMPAC-original: I have a pet dog.
COMPAC-revised: I think you really need to go shopping.
COMPAC-paraphrase: I like shopping.
COMPAC: Yes, I prefer to be fashionable.

Table 9: Sample generations by different models.

Persona: I just want to have fun with my friends. I don't drink or do drugs or anything. I am 19 and I cannot wait to move out of my parents home.
User: Are you enjoying life ?
GPT2 (2019): Yes! What about you?
COMPAC-original: Hey yes! I don't drink.
COMPAC-revised: Yes. I like to go out with friends.
COMPAC-paraphrase: I don't want to be alone.
COMPAC: Not really. I want to have a party.

Table 10: Sample generations by different models.

Persona: I've rainbow hair. I am an animal activist. I spend my time bird watching with my cats.
User: What motivates you to work with animals?
GPT2 (2019): I love bird-watching.
COMPAC-original: I love bird-watching. What about you?
COMPAC-revised: I love animals.
COMPAC-paraphrase: I am in love with animals.
COMPAC: I want to protect animals to make a difference in this world.

Table 11: Sample generations by different models.

Persona: I'm always wearing a hat. I've a broken arm. My friends are all skateboarders.
User: Do you skate? I do.
GPT2 (2019): Yes. How about you?
COMPAC-original: I wear a hat.
COMPAC-revised: Yes. My friends are skateboarders.
COMPAC-paraphrase: That's great. My friends are skateboarders.
COMPAC: My friends and I go to the park for skateboarding.

Table 12: Sample generations by different models.

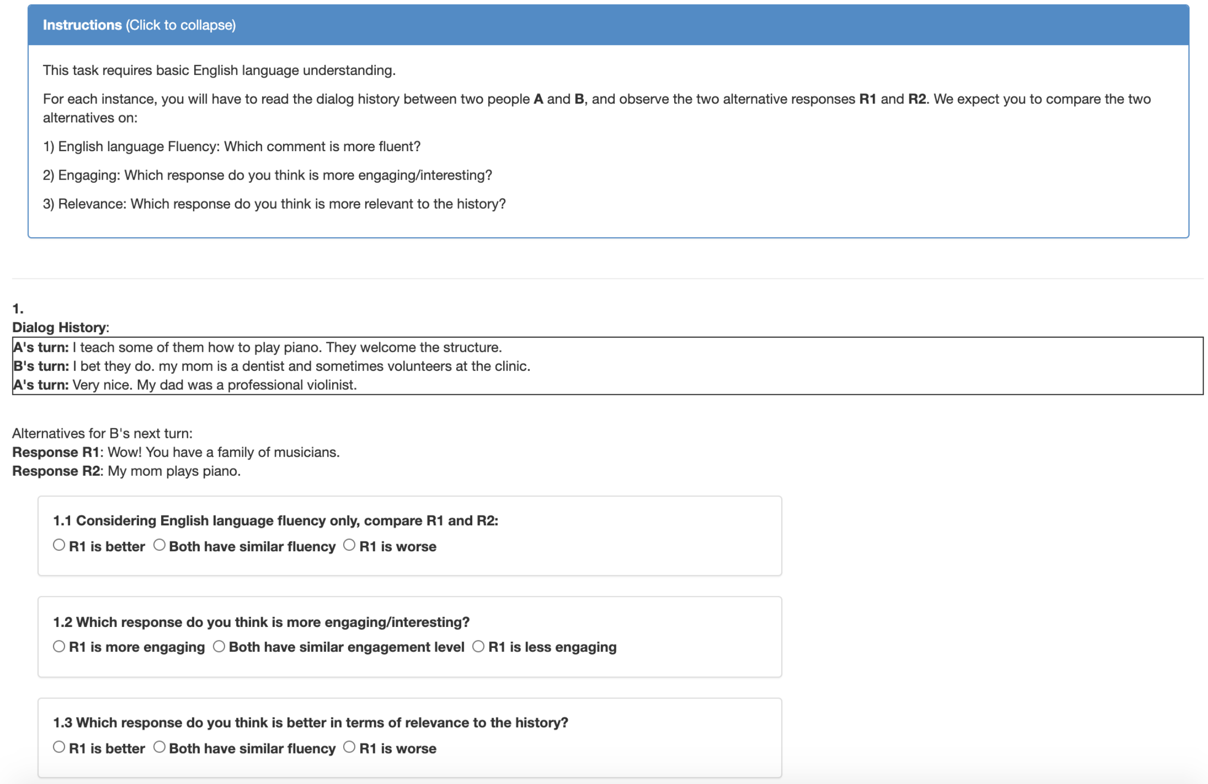

Figure 4: Human evaluation for pairwise comparison between COMPAC and another baseline.

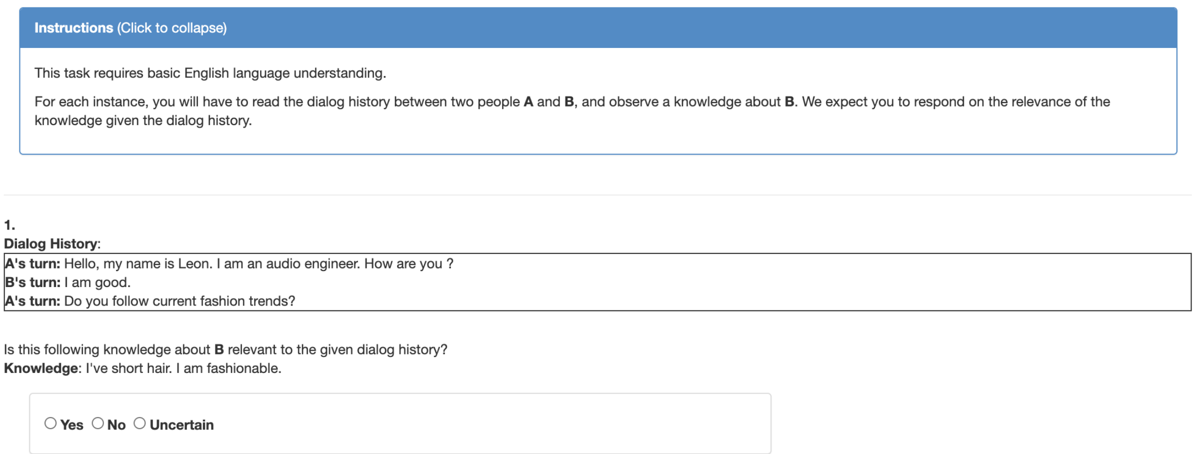

Figure 5: Human evaluation for assessment of persona grounding performance with expansions.