Rethinking Personalized Natural Language Generation with the PersonaSocialNorms Corpus and Ranking Evaluation

Anonymous ACL submission

Abstract

Personalized language generation is playing an increasingly significant role in language technologies. Persona-based generation is a personalization approach that conditions the generation of descriptive sentences about an individual and has been shown to successfully emulate the language characteristic of individuals with these traits. This is a challenging task to design, model, and evaluate, and as such, early work in this area approached the problem with constraints to simplify the problem. We argue that the way forward requires modifications to these restrictions in three key areas; (1) realistic conversational data, (2) representative and diverse persona sentences, and (3) modified ranking evaluation. We present an extension of the Social-Chem-101 corpus, the PersonaSocialNorms corpus, which contains a collection of Reddit posts about social situations and written judgements from others stating that the actions taken by the original poster are right or wrong. Our corpus contains a collection of 95K judgements written by 6K authors filtered from the Social-Chem-101 corpus. We extend the data with 20-500 persona sentences for each author. By using more realistic data, we find previous persona consistency metrics inadequate for evaluation. We provide a novel ranking evaluation and implement several architectures inspired by recent work, showing promising results and room for improvement.

1 Introduction

Personalization is of growing importance in natural language technologies as users expect systems to cater to their specific needs. In particular, there is a growing interest in a perspectivist approach to many natural language processing (NLP) tasks, which emphasizes that there is no single ground truth (Aroyo and Welty, 2015; Basile et al., 2021). This is a more common view in generation tasks, as it is easier to see that multiple translations or continuations of a dialog are correct. However, work in this area tends to not take additional contextual factors into account during generation. Flek (2020) emphasized the need to interpret language with it's personal contextual factors to create higher performing personalized systems. Dudy et al. (2021) similarly argue that additional contextual information should be incorporated in such models, particularly for generation.

Work on personalized or persona-based dialog systems has begun to incorporate contextual information in response generation. The work of Zhang et al. (2018) introduced the PersonaChat dataset, where two crowd workers converse with each other while attempting to emulate a persona described by five short sentences. Models developed using this data condition on encoded persona sentences. Dinan et al. (2020) extended this dataset with rephrasings of the utterances to avoid high direct word overlap with persona sentences, yet these dialogs focus directly on incorporating information from a few short phrases. Workers were instructed to use these facts in their conversations, which leads to artifacts, such as the unprompted addition of personal information to the end of unrelated utterances (e.g. “I am a lifeguard” in response to someone saying they will read a book). They do not accurately reflect the real world, e.g. "to stay in shape, I chase cheetahs at the zoo”, and they ask people to emulate an identity whose life experiences (e.g. getting divorced, living in different places, being a lawyer, owning a business) could plausibly shape their views of interpersonal conflict described in our data, but through the shallow nature of crowdsourced conversations and lack of real lived experience of participants, fails to be reflected in the PersonaChat dialogs.

An example from our PersonaSocialNorms corpus can be found in Figure 1. We see a user asking if they did something wrong in a conversation with their girlfriend about whether or not to terminate a pregnancy. There are two responses from other users with different judgments of the situation (NTA = not the asshole, YTA = you are the asshole). On the left, we see persona sentences for each user. One user appears to be more family-oriented than the other which may impact their judgement of the situation. In our initial human evaluation, we found that generated or human responses were always rated as consistent with a given set of persona sentences in this corpus, as opposed to work on PersonaChat where consistency is more directly related to the incorporation of facts about oneself. Which one more closely matches a given persona only becomes clearer when we compare multiple responses. Although a few other works have evaluated the ranking of generated responses, we add crucial comparisons to the ranking and note that it's importance for evaluating realistic personalized response generation has not been emphasized.

We argue that future work can improve persona-based generation models with three modifications to their approach. The first is the use of realistic data. Our corpus contains 95K judgements of social situations written by 6K authors filtered from Social-Chem-101. Second, we suggest that models benefit from having a larger pool of persona sentences that are written by the same person who writes the judgements, and we crawl 20-500 persona sentences per author. Author responses contain judgements of social situations that require a deeper understanding of personal context than casual open dialog used in previous work. We develop several architectures inspired by recent work for persona-based generation, finding that our FlanT5 Twin Encoder with similar persona sentences outperforms other models. Furthermore, we find that by training a model for generating user judgements, we also score competitively with previous data perspectivist work on judgement prediction, even outperforming their models in one setup. Third, we find the consistency evaluation insufficient when using more realistic data and suggest a ranking evaluation. We will release our corpus, code, and human evaluations.

2 Related Work

Personalized Datasets One of the earliest areas with a focus on personalization has been recommender systems, where personalization is an important part of large-scale industry systems (Davidson et al., 2010; Konstan and Terveen, 2021; Xu et al., 2022). Personalized dialog generation is an-

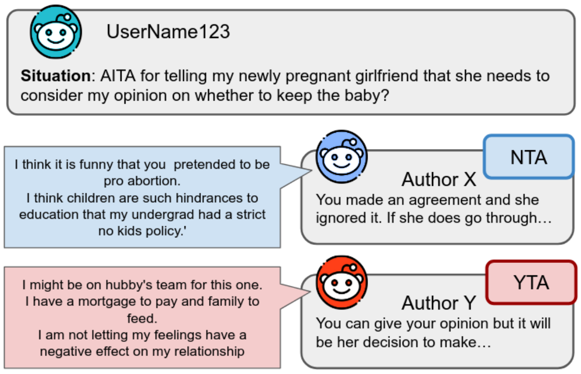

Figure 1: Example of a post in AITA subreddit. The example includes a situation title and two comments with different perspectives regarding the situation, plus persona sentences for the respective users.

other field where the use of persona sentences has been extensively explored. There have been several datasets that focus on building persona-based dialog generation models using social media sources like Reddit (Al-Rfou et al., 2016; Wu et al., 2021), Twitter (Li et al., 2016b), and Weibo (Zheng et al., 2019) where for each speaker there are five personality traits rather than sentences.

Zhang et al. (2018) introduced Persona-Chat dataset with 1k crowdsourced personas. Mazaré et al. (2018) introduced an approach to extract persona sentences from Reddit by pattern matching. Zhong et al. (2020) collected conversations and persona sentences from Reddit for the purpose of generating empathetic dialog. They use up to 10 persona sentences extracted randomly for their experiments. These two works are the similar to ours in their construction of more realistic corpora, however, they focus on the task of response selection rather than generation.

Meanwhile, work on generation has used automatic and human-evaluated consistency metrics (Madotto et al., 2019), which ask if utterances are entailed by a persona or how well utterances match persona sentences on a numerical scale. While this may work well for more artificial datasets, for example where an utterance says "I am about to watch Game of Thrones" and a persona sentence says "I love watching game of thrones", we find that more realistic scenarios are not as straightforward. Our dataset is instead constructed from the profiles of real people who wrote both the judgements of social situations and their persona sentences.

Additionally, several works have introduced datasets for personalized language generation for various tasks. Majumder et al. (2019) introduced a new task of personalized recipe generation. Vincent et al. (2023) released a dataset that contains movie dialogs conditioned on character descriptions. Joshi et al. (2017) extended the bAbI dialog dataset with user profile information. Yessenalina et al. (2010) looked at generating rationales for sentiment analysis, finding that they improved prediction performance. Recently, Salemi et al. (2023), introduced a novel benchmark for training and evaluating language models for personalized text classification and generation.

Personalized Models Personalized generation models, attempt to generate a response given an input utterance and additional personal contextual information. Li et al. (2016b) introduce a speaker model that models only the speaker and an extension speaker-addressee model which models both the speaker and addressee. Madotto et al. (2019) use only a few dialog samples to generate personalized responses, by casting personalized dialog learning as a meta-learning problem. Moreover, other works, have modified sequence-to-sequence frameworks to infuse persona information in the decoder (Zheng et al., 2019), or in the transformer framework by adding an attention routing mechanism that controls the contribution of persona sentences in the decoding process (Zheng et al., 2020). Extending sequence-to-sequence networks with memory networks is a common approach to infusing persona information. Song et al. (2019) introduce Persona-CVAE, which is a memory-augmented architecture that aims to exploit the persona information from the given context and also generate diverse responses. Ma et al. (2021) introduced DHAP, which consists of a history encoder, personalized post encoder, user history memory, and personalized decoder to fuse the learned user profile into the response generation process. Wu et al. (2021) propose a generative split memory network, to use information from a user profile memory network, and a comment history memory network. Recently, Soni et al. (2022) introduced HaRT, a large-scale transformer model which contains a user-state attention layer. They apply the model to several downstream tasks like stance prediction and demographic inference. Recently, Huang et al. (2023) introduced the Persona-Adaptive Attention (PAA) model. The PAA model combines two encoders to encode the dialog context and persona sentences, with persona-adaptive attention in the decoding layer.

3 Dataset

We used the dataset of Welch et al. (2022b) as the foundation of our work. The authors collected data from Reddit, an online platform with many separate, focused communities called subreddits. The data is from the AITA subreddit, where users share descriptions of social situations that they are involved in and ask members of the community for their opinions. These members assess if the poster is the wrongdoer in the described situation. They provide a verdict in the form of "you're the asshole" (YTA) or "not the asshole" (NTA). The dataset was filtered from Forbes et al. (2020)'s Social-Chem-101 corpus but also includes the post title, full text, all comments, and their corresponding authors. We refer to the post title as the *situation*, as the title is usually a short description of the conflict situation. The comments are preprocessed in order to extract those that contain a verdict of YTA or NTA,[1] and others were removed. In order to extract verdicts, they manually created a set of keywords for both classes and filtered the comments to remove these expressions. The initial dataset contains 21K posts, and 364K verdicts (254K NTA, 110K YTA) written by 104K different authors.

3.1 Persona Extraction

Furthermore, we expand the dataset by retrieving the comment histories for each user in the dataset. To extract the persona sentences for the users, we adopt the approach described in Mazaré et al. (2018). Initially, we split each comment into a sentence and kept only sentences that contain between 5 and 20 tokens. Then we add two constraints to each sentence in order to classify it as a persona sentence; (1) it must contain the tokens *I*, *my* or *mine* and (2) one verb, one noun, and one pronoun or adjective.

After performing these steps, we obtained a set of persona sentences for each user. Additionally, we filtered our dataset to include only those users who contain more than 20 persona sentences and less than 500 persona sentences. Our final dataset contains 20K posts and 95K verdicts written by 6K different authors, which we will release upon publication as the PersonaSocialNorms corpus.

[1] Reddit posts were crawled with the Reddit API (`https://www.reddit.com/dev/api`) and comments with the PushShift API (`https://files.pushshift.io/reddit/comments/`).

3.2 Comparison to PersonaChat

In an effort to quantify the differences between PersonaChat and our corpus, we measured the unigram and bigram Jaccard similarity between persona sentences and author responses. We calculated the maximum similarity between any persona sentence for an individual and their given response. This follows the idea that PersonaChat directly incorporates facts from the persona, leading to high similarity between a persona sentence and a given dialog response. We report this value averaged across all users for each corpus. We found the unigram similarities to be 0.16 and 0.12 for PersonaChat and our coprus, respectively. Our corpus had a max bigram similarity of 0.01, whereas PersonaChat's was four times higher at 0.04. This shows that even after efforts were made to reduce direct overlap in the PersonaChat corpus (also known as ConvAI2), the similarity between the persona sentences and responses is high.

4 Problem Formulation

Our task considers as a data point, a post that contains a summary of the situation description, a comment of the post containing a personal verdict about the situation, and the author of the verdict jointly with the corresponding persona sentences. Therefore, for our generation task, we have three components: (i) the input sequence which corresponds to the main post, (ii) the target output sequence which corresponds to the comment containing the verdict, and (iii) the user's persona sentences. For a given situation post s written from a random author a, we have a set of comments $C_s = \{c^s_{a_1}, c^s_{a_2}, \ldots, c^s_{a_n}\}$ written by n different authors. Each post describing a situation s contains many comments $c^s_{a_i} \in C_s$, and an author a has many comments $c^{s_i}_a$ on different posts s_i. Hence, as we have different target outputs, for the same input sequence, we need additional information to condition our model. The generation task can be formalized as $p(c^s_a|s, a)$. For each author a the model can take advantage of $P_a = \{p^a_1, p^a_2, \ldots, p^a_k\}$, where p^a_i, denotes the $i-$th persona sentence for author a. We describe two different methods to extract a set of k persona sentences for each user in the dataset.

Random sampling In this setup, we randomly sample up to k persona sentences for each user.

Most relevant sampling We compute embeddings using SBERT (Reimers and Gurevych, 2019), for all extracted persona sentences and situation titles

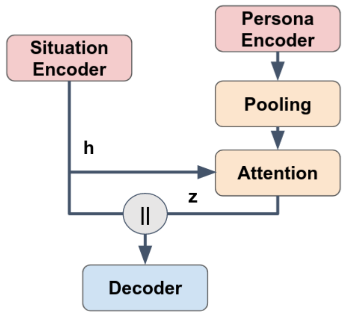

Figure 2: Twin encoder model, with an extra encoder to model the auxiliary user information.

in our dataset. We compute the cosine similarity between an author's persona sentences and the situations that they have commented on and select the top k most similar persona sentences for each situation. We aggregate the top k across situations for each author and rank the persona sentences by their frequency, again keeping the top k.

5 Methodology

After discussing the base transformer, we describe two modifications to the encoder-decoder architecture in order to incorporate additional information.

5.1 Base Transformer

The main architecture used in our models is an encoder-decoder transformer model (Vaswani et al., 2017). The architecture aims to model $p(y|x)$. The encoder takes as an input a sequence $\mathbf{x} = \{x_1, \ldots, x_n\}$ and maps it into a sequence of representations $\mathbf{h} = \{h_1, \ldots, h_n\}$. Given $\mathbf{h}$, the decoder generates an output sequence $\mathbf{y} = \{y_1, \ldots, y_m\}$.

Given the input sequence $s = [w_1, \ldots, w_{n_s}]$, we utilize a pre-trained transformer encoder to embed the tokens of the sequence $h = encoder(s; \theta^{(enc)})$, where $h \in \mathcal{R}^{d \times n_s}$ where d is the output dimension of the encoder and n_s is the size of the input sequence. In general, in the transformer, the output probabilities can be computed as:

$$\begin{aligned} o &= decoder(h; \theta^{(dec)}) \\ \hat{y} &= softmax(\mathbf{W_o}^\top o) \end{aligned} \quad (1)$$

where $\mathbf{W_o} \in \mathcal{R}^{d \times v}$ is the language model head where v is equal to the vocabulary size, and $o \in \mathcal{R}^{d \times n_t}$, are the last decoder state for the output sequence, where n_t is the size of the target sequence.

5.2 Twin Encoder

In Figure 2, we show the architecture of our first model Twin Encoder. As we described in §4, we

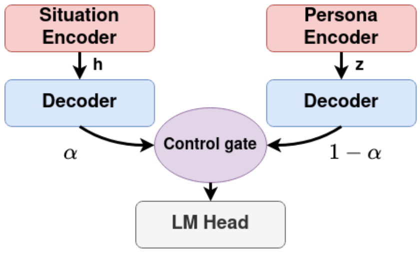

Figure 3: Style decoder model, with a decoder that focuses on persona style, and a control gate that controls the amount of information used from both decoders.

are attempting to model $p(c_a^s|s,a)$, where s is the input sequence, c_a^s is the target output and a is the additional information. The sequence of persona sentences is given by $a = [p_1^a, \ldots, p_{m_a}^a]$, where $a \in \mathcal{R}^{m_a \times n_p}$. m_a is the number of persona sentences, and n_p is the maximum token length in the persona sentences. We utilize a pre-trained transformer encoder to compute a final representation as $z = pool(encoder(a; \theta^{(enc)}))$, where $z \in \mathcal{R}^{d \times m_a}$, and $pool(\cdot)$, performs a mean-pooling over the tokens of each persona sentence. Furthermore, we compute a final representation of the auxiliary information as $\bar{z} = Att(h, z)$, where $\bar{z} \in \mathcal{R}^{d \times n_s}$ $Att(\cdot)$ is an attention layer as in (Vaswani et al., 2017) where the representation h of the input sequence is the query and z is the key and value. Then, we compute the decoder state as $o = decoder(\mathbf{W}^c[h||\bar{z}]; \theta^{(dec)})$ where $\mathbf{W}^c \in \mathcal{R}^{d \times 2d}$, and $||$ is the concatenation operator.

Our twin encoder (TE) architecture is similar to the PAA model introduced in previous work (Huang et al., 2023). Both models employ two encoder layers to model both the input context and the persona. However, the key distinction between these models lies in their approach to information processing within the decoder. The PAA model performs two cross-attentions over both encoders in the decoder and then combines the information afterward, while the TE architecture combines the encoder's information beforehand and subsequently performs one cross-attention in the decoder.

5.3 Style Decoder

In the second modification (Figure 3), we concatenate all auxiliary sentences to create the sequence of tokens $a = [w_1^{a,1}, \ldots, w_{n_p}^{a,1}, \ldots, w_1^{a,m_a}, \ldots, w_{n_p}^{a,m_a}]$. We utilize a pre-trained transformer encoder to compute the representations, $z = encoder(a; \theta^{(enc)})$ where $z \in \mathcal{R}^{d \times n_p m_a}$. Afterward, we compute the output distribution $\hat{y}$ as follows:

$$\begin{aligned} o' &= decoder(z; \theta^{(dec')}), \\ \hat{y} &= softmax(\mathbf{W_o}^\top(\alpha \cdot o + (1-\alpha) \cdot o')) \end{aligned} \tag{2}$$

where $o' \in \mathcal{R}^{d \times n_t}$ are the writing style decoder states, and $\alpha \in \mathcal{R}^{n_t}$. α is a learnable parameter and contains a scalar in the range of [0,1], that controls the amount of information to use out of different language heads. We compute $\alpha = \sigma(\mathbf{V}(\mathbf{W_c}[o||o'])$ where $\mathbf{W_c} \in \mathcal{R}^{d \times 2d}$, $\mathbf{V} \in \mathcal{R}^{d}$, and $\sigma(\cdot)$ is the sigmoid function. From the equation, the computation of α is similar to the gate computation in (Chung et al., 2014), with similar approaches used in previous works to fuse stylistic information during generation (Zhou et al., 2018; Zheng et al., 2019).

6 Experiments

In our experiments, we utilize two base models, that follow an encoder-decoder architecture. To incorporate personalization, we are using two different methods during training that add user information in the encoder and do not change the architecture of the models:

Priming. This method was originally used in recurrent neural networks. It initially passes information about a user through the model, and then the text that needs to be classified (King and Cook, 2020). In our approach, we sample a number of sentences from a user's history that are up to a maximum number of m tokens in order to fit into the context window of the model. Then, we concatenate this sampled text for each user at the beginning of the input text for the encoder during training.

User ID. In this approach, we append a special user token, at the end of the input text for the encoder during training. Several methods incorporate the user ID to learn user representations in the model (Li et al., 2016b; Welch et al., 2022a). However, one drawback of this method is that it cannot generalize to unseen users during test time.

We also adapt the recent PAA model (Huang et al., 2023), which has shown superior performance on the PersonaChat task, to run on our dataset and compare with our proposed architectures. For the PAA model, we utilize only the persona sentences as an auxiliary input. We are using the modified architectures, (§5) twin encoder (TE) and style decoder (SD), with two different types of auxiliary information for each user; (1) persona sentences (PS). These sentences are extracted using the methods described in §3.1, and (2) comments (C), which are other comments from the user in the AITA subreddit.

6.1 Zero/Few Shot Learning

In addition to fine-tuning, we explore zero and few-shot learning by utilizing large transformer models that contain billions of parameters, making them around 100 larger than our models. In the zero-shot setup, we adjust the prompt in order to include up to 10 examples of auxiliary information (either persona sentences or comments). On the other hand, in few-shot learning, we only utilize pairs of past situation titles and comments of an author to construct the prompts for the models.

6.2 Perspective Classification

We also evaluated our model on the perspective classification task from previous work by extracting the labels (NTA/YTA) from the generated comments. We use the three splits from Plepi et al. (2022). The first split is the verdict split, which is our default split for all experiments. Additionally, we perform situation and author splits, which have disjoint sets of situations and authors respectively, across train, validation, and test. We experiment with our two top-performing models, finding that our models are competitive and outperform previous work on the situation split (see B).

6.3 Experimental Setup

We train our models for 10 epochs, with the AdamW optimizer, using an initial learning rate of $5e-5$. We use a linear learning rate scheduler with 100 warm-up steps and early stopping on the validation set. As our base models, we are using BART (Lewis et al., 2020) and FlanT5-base (Chung et al., 2022), with a maximum input length of 512, and a maximum target length of 128. BART models have up to 180M parameters, while FlanT5 models go up to 320M. For the twin encoder architectures, we found that encoding the persona separately leads to better performance, while for the style decoder, the persona sentences are concatenated to create a long context. For the zero/few-shot learning, we use the XXL model of Flan-T5, with 11B parameters. We experimented with the optimal number of persona sentences, finding that $k = 20$ performed best (see Appendix A). In the priming method, we sample $m = 100$. Our experiments run on a single NVIDIA A100 40GB GPU with an average running time (training + inference) of 6 hours. For the PAA model, we use the GPT2-medium to initialize the decoder and keep the configurations the same as described in (Huang et al., 2023). The PAA model has 475M parameters.

6.4 Evaluation metrics

Automatic Evaluation In the automatic evaluation for the generation task, we utilize two-word overlap-based metrics: BLEU (Papineni et al., 2002) and ROUGE (Lin and Och, 2004). BLEU evaluates the quality of generated text by computing the n-grams overlap with the original comment. ROUGE is a recall-oriented adaptation of the BLEU. Instead of using n-grams, ROUGE uses the longest common subsequence to compute the F1 score. Moreover, we also use the diversity metric, to compute the number of distinct n-grams generated by the model (Li et al., 2016a). In addition, we also compute DistS-n, which is the average number of distinct tokens across situations. Computed perplexities were in the range of 15-25, but these do not reliably indicate performance as the vocabularies for BART and FlanT5 are different.

Human Evaluation In addition to automatic metrics, we also perform a human evaluation using Prolific [2]. Due to the costs of human evaluation, we only performed a human evaluation for our top two models, FlanT5 + TE (PS), BART + TE (PS), and FlanT5 + SD (C) which was the highest-performing style decoder model. We randomly sample 100 examples from the test set and conduct our human evaluations in two parts. In the first part, we focus on persona matching with the generated comments.

Our initial human evaluation was similar to that of prior work which measured persona consistency. Annotators were asked if a response was consistent with a persona when presented with 20 persona sentences. We found that in almost every case the answer was yes. This evaluation is insufficient for the PersonaSocialNorms corpus where it is unlikely for persona sentences to be directly stated or even rephrased in someone's comments.

Instead, we developed a ranking evaluation. Others have used a ranking of models as an evaluation, but have not ranked the response with human responses (Song et al., 2019; Tang et al., 2023). In our novel setup, we show the annotators a set of $k = 20$ most relevant persona sentences from a

[2] We paid 12$ per hour of annotations.

Model	**BLEU-1** ↑	**BLEU-2** ↑	**R-1** ↑	**R-L** ↑	**Dist-1** ↑	**Dist-2** ↑	**DistS-1** ↑	**DistS-2** ↑
PAA (Huang et al., 2023)	15.0	5.1	18.9	16.3	0.01	0.06	0.41	0.53
BART + Priming	4.6	1.9	18.4	14.8	0.02	0.14	0.52	0.61
BART + User Id	4.1	1.7	18.7	15.2	0.03	0.15	0.54	0.63
BART + TE (PS)	9.9	4.2	25.4	**19.7**	0.033	0.17	0.5	0.57
BART + TE (C)	5.0	2.45	18.8	15.6	0.029	0.14	0.52	0.62
BART + SD (PS)	4.2	2.0	19.1	15.8	0.03	0.15	0.41	0.55
BART + SD (C)	5.8	2.45	23.5	18.8	0.03	0.16	0.47	0.63
FlanT5 + Priming	10.7	4.2	15.7	13.6	0.02	0.1	0.59	0.75
FlanT5 + User Id	5.7	2.4	19.9	15.7	0.029	0.14	0.61	0.77
FlanT5 + TE (PS)	**25.3**	**9.0**	**25.6**	17.6	**0.053**	**0.387**	**0.73**	**0.92**
FlanT5 + TE (C)	7.6	2.9	18.2	12.0	0.032	0.25	0.62	0.73
FlanT5 + SD (PS)	11.9	5.1	17.1	11.4	0.04	0.29	0.65	0.8
FlanT5 + SD (C)	18.3	5.9	18.8	12.5	0.04	0.29	0.64	0.79

Table 1: Automatic metrics of fine-tuned models, for our based models with priming, user id, twin encoder (TE), and style decoder (SD). We report BLEU-1, BLEU-2, ROUGE-1 (R-1), ROUGE-L (R-L) scores in the range of 0-100 and diversity metrics in the range 0-1. (PS) means the model uses persona sentences as additional information, (C) past comments. The auxiliary set of information is extracted using the most similar method.

Model	**BLEU-1**	**BLEU-2**	**R-1**	**R-L**
XXL ZS (PS)	6.2	1.6	11.2	7.4
XXL ZS (C)	2.5	0.7	10.4	7.1
XXL FS	**11.7**	**3.9**	**15.8**	**11.6**
Base ZS (PS)	0.84	0.3	7.6	5.2
Base ZS (C)	0.67	0.24	7.4	5.0
Base FS	2.8	0.63	8.2	6.4

Table 2: Automatic metrics (R=ROUGE) of zero-shot (ZS) and few-shot (FS) learning of FlanT5-XXL with 11B parameters and FlanT5-base with 250M.

Model	**Generated over Incorrect** ↑	**Generated over Correct** ↑
BART + TE (PS)	62.8%	38.9%
FlanT5 + TE (PS)	**67.2%**	**42%**
FlanT5 + SD (C)	49.4%	39.4%

Table 3: Human evaluation results related to the ranking of comments with respect to the given persona. Correct is ranked over incorrect 70.8% of the time, providing an upper bound for generated over correct.

Model	**Fluency** ↑	**Relevance** ↑
BART + TE (PS)	**43%**	**42%**
FlanT5 + TE (PS)	30.6%	25.6%
FlanT5 + SD (C)	41.7%	40%

Table 4: Human evaluation results for our top two models BART and FlanT5 fine-tuned with Twin Encoder (TE) with persona sentences (PS), and FlanT5 + Style Decoder (SD), with comments.

user a, and three comments: the comment of author c_a^s, the generated comment from the model for that user, and a comment $c_{a'}^s$, written by another user a', for the same situation s. Then we ask the annotators to rank the comments with respect to the "possibility that they have been written by the user with the given persona sentences." Ranking with both correct and incorrect human responses allows us to more clearly understand model performance. It is more difficult for models to be ranked over the ground truth than it is to outperform other generated responses. We find that 70.8% of rankings have the correct human response over the incorrect one. This gives us an upper bound on model performance.

In the second part of our evaluation, we focus on the fluency and relevance of the comment with respect to the situation. We show annotators the situation summary title s, and two comments: the gold comment c_a^s, and the corresponding generated comment from our model. We ask the annotators to pick the most fluent comment and the most relevant comment with regard to the given situation summary.

7 Results and Analysis

Extraction method In Table 1, we report the automatic results for all combinations of architectures from our models. In general, the FlanT5 variations proved to perform better, which may be attributed to the size difference of the base models (250M vs 140M). Furthermore, BART-based models were the most sensitive with respect to the retrieval method used to extract the set of persona sentences or comments. When random persona sentences and comments were utilized, the generation of the BART-based model would degrade, and upon manual inspection of the results, the generated output would contain only "NTA/YTA" tokens.

Architecture Comparison The best-performing architecture across both models is the twin encoder. The key difference between the two architectures is that information about the situation and the auxiliary context is combined. In the twin encoder architecture, information is combined before the decoder performs the cross-attention with the encoder states, while in the style decoder, the information is combined after the decoder. Hence, in our case, it proved to be more useful to use only one decoder layer and combine the information earlier, as opposed to previous work (Zheng et al., 2019). In addition, FlanT5 + TE (PS) performs better than the PAA model despite having fewer parameters. Moreover, FlanT5 + TE (PS), has the most diverse responses, even across situations, with scores close to the original responses on Reddit[3]. Among priming and user ID, that do not require any architecture changes, priming proved to be better. However, in the case of FlanT5 + priming, it generated excessively long responses resulting in nonsense judgments.

Zero/Few Shot Learning Table 2 shows the results of zero and few shot learning for FlanT5-base and FlanT5-XXL. Overall FlanT5-XXL showed better zero/few shot performance, which indicates that larger models are better in context learning (Brown et al., 2020). Zero-shot learning proved more difficult. However, for few-shot learning, FlanT5-XXL is better and comparable to the results of some of our fine-tuned models. Nevertheless, it is performing worse than our top two models, despite having almost 100 times more parameters.

Human Evaluation In Table 3 we show the results for the first part of the survey, which is related more to alignment between the generated response and the persona of the user. We report the average accuracy for the number of times the generated comment was higher in rank over the incorrect and the correct one. FlanT5 + TE (PS), is performing the best across all metrics, with almost 5% better accuracy in selecting the generated comment over the incorrect one. This finding suggests that the more diverse responses align closer to the persona sentences of the users [4]. The agreement between annotators is 0.45 for the FlanT5 + TE (PS), which is a moderate agreement, while the other two models show fair agreement with 0.27 and 0.22. The results for the human evaluation related to comment fluency and relevance, are shown in Table 4. We report the average accuracy of human annotators in selecting the generated comment in the evaluation. Human annotators selected the BART + TE (PS) model most often. The main reason for these results might be due to the length of the comment. BART + TE (PS), on average, has shorter responses (25.3 for BART versus 49.9 for FlanT5). The Cohen Kappa for these annotations is 0.3 for FlanT5 + TE (PS), 0.27 for BART + TE (PS), and 0.24 for FlanT5 + SD (C), which shows a fair agreement between the annotators.

8 Conclusions

As we make progress in the area of natural language generation, we will need to have models that take additional contextual information into account, especially personal contextual factors. We discussed the limitations of previous work on persona-based dialog and three areas of improvement. First, we investigated the differences between artificial and realistic personas and introduced the PersonaSocialNorms corpus, which contains real personas and judgements of conflict situations. Second, we encouraged the use of representative and diverse persona sentences. Our corpus contains 20-500 persona sentences per author, more than previously released corpora. Persona sentences are written by the same person as the response. We experimented with ways to incorporate the persona information, finding using sentences most similar to the situation worked best. Third, we found that previous consistency evaluation metrics were inadequate when using our corpus and suggested a novel ranking human evaluation. We also implemented two novel architectures inspired by recent work, finding that our FlanT5 twin encoder model outperformed our style decoder approach and recent work in this area. Additionally, we found that our generation model performed competitively with previous work on perspective classification. We will release our code and corpus upon publication.

[3]DistS-1 and DistS-2 for original comments on Reddit were 0.76 and 0.93 respectively.

[4]Examples of the generated comments are in Appendix C.

Limitations

In this work, we utilize persona sentences extracted from Reddit in order to improve personalized judgment generation in social media. However, there are a lot of persona sentences available per user. Even though we attempted to sample the most relevant persona subset for each user, some of those might not be as useful, and future work can explore other methods to have more control over the quality of personas extracted. Moreover, in this work, we train and modify only base models, instead of large ones, due to computation resources. We attempt to utilize the large models (FlanT5-XXL), by performing zero/few shot learning, however, we do not try to fine-tuning those.

Performing human evaluation using the persona sentences, has high costs due to the considerable amount of information that the annotators need to evaluate in order to decide if a comment matches the given persona. Therefore, we only performed human evaluation in our top-performing models with automatic metrics. In future work, it might be useful to increase the number of evaluated models, by lowering the costs of human evaluation with the improved quality and quantity of extracted persona sentences.

Ethical Considerations

Personalized models use the personal information of users on social media in order to improve performance. However, this requires us to address a range of ethical considerations related to our work, like privacy and consent, bias, and responsible use of the technology. The use of personalization data will be transparent, and anonymized (Hewson and Buchanan, 2013). Language generation with personalized information can enhance the automatic generation of perspectives, opinions, or stances in social media. While this might be helpful in some NLP applications, it might be undesired and harmful in some other cases. Researchers should take into account users' expectations when using and collecting data from social media (Townsend and Wallace, 2016; Williams et al., 2017).

Moreover, bias in the model can cause misinterpretation or negatively influence different communities (Blodgett et al., 2020). The underrepresented communities in our data, may be affected negatively by the usage of personalized models. Hence, we suggest that the users should be aware of how their data is being used, and given the choice of not using their data from training such personalized models.

References

Rami Al-Rfou, Marc Pickett, Javier Snaider, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2016. Conversational contextual cues: The case of personalization and history for response ranking. *arXiv preprint arXiv:1606.00372*.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. Toward a perspectivist turn in ground truthing for predictive computing. *arXiv preprint arXiv:2109.04270*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. 2010. The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296.

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

Shiran Dudy, Steven Bedrick, and Bonnie Webber. 2021. Refocusing on relevance: Personalization in nlg. *arXiv preprint arXiv:2109.05140*.

Lucie Flek. 2020. Returning the N to NLP: Towards contextually personalized classification models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7828–7838, Online. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online.

Claire Hewson and Tom Buchanan. 2013. Ethics guidelines for internet-mediated research. The British Psychological Society.

Qiushi Huang, Yu Zhang, Tom Ko, Xubo Liu, Bo Wu, Wenwu Wang, and H Tang. 2023. Personalized dialogue generation with persona-adaptive attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12916–12923.

Chaitanya K Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in goal-oriented dialog. *arXiv preprint arXiv:1706.07503*.

Milton King and Paul Cook. 2020. Evaluating approaches to personalizing language models. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2461–2469, Marseille, France. European Language Resources Association.

Joseph Konstan and Loren Terveen. 2021. Human-centered recommender systems: Origins, advances, challenges, and opportunities. *AI Magazine*, 42(3):31–42.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612.

Zhengyi Ma, Zhicheng Dou, Yutao Zhu, Hanxun Zhong, and Ji-Rong Wen. 2021. One chatbot per person: Creating personalized chatbots based on implicit user profiles. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 555–564, New York, NY, USA. Association for Computing Machinery.

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

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2019. Generating personalized recipes from historical user preferences. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5976–5982, Hong Kong, China. Association for Computational Linguistics.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022. Unifying data perspectivism and personalization: An application to social norms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7391–7402, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*.

Haoyu Song, Wei-Nan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting persona information for diverse generation of conversational responses. *arXiv preprint arXiv:1905.12188*.

Nikita Soni, Matthew Matero, Niranjan Balasubramanian, and H Andrew Schwartz. 2022. Human language modeling. *arXiv preprint arXiv:2205.05128*.

Yihong Tang, Bo Wang, Miao Fang, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2023. Enhancing personalized dialogue generation with contrastive latent variables: Combining sparse and dense persona. *arXiv preprint arXiv:2305.11482*.

Leanne Townsend and Claire Wallace. 2016. Social media research: A guide to ethics. *University of Aberdeen*, 1:16.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Sebastian Vincent, Rowanne Sumner, Alice Dowek, Charlotte Blundell, Emily Preston, Chris Bayliss, Chris Oakley, and Carolina Scarton. 2023. Personalised language modelling of screen characters using rich metadata annotations. *arXiv preprint arXiv:2303.16618*.

Charles Welch, Chenxi Gu, Jonathan K. Kummerfeld, Veronica Perez-Rosas, and Rada Mihalcea. 2022a. Leveraging similar users for personalized language modeling with limited data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1742–1752, Dublin, Ireland. Association for Computational Linguistics.

Charles Welch, Joan Plepi, Béla Neuendorf, and Lucie Flek. 2022b. Understanding interpersonal conflict types and their impact on perception classification. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science*.

Matthew L Williams, Pete Burnap, and Luke Sloan. 2017. Towards an ethical framework for publishing twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, 51(6):1149–1168.

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

Jiajing Xu, Andrew Zhai, and Charles Rosenberg. 2022. Rethinking personalized ranking at pinterest: An end-to-end approach. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 502–505.

Ainur Yessenalina, Yejin Choi, and Claire Cardie. 2010. Automatically generating annotator rationales to improve sentiment classification. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 336–341, Uppsala, Sweden. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*.

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9693–9700.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566, Online. Association for Computational Linguistics.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

A Analysis of Persona Context Size

We report in Table 5, the results for FlanT5 + TE (PS), with different amounts of persona sentences as context. Our experiments are run with persona amounts $\{5, 10, 15, 20\}$. We notice that the best-performing model is using 20 persona sentences. However, the differences between the models' performance are small, and one can trade off small performance values, with computational speed-up, by using only the top-5 persona sentences.

Sentences	**BLEU-1**	**BLEU-2**	**R-1**	**R-L**
5	24.1	8.4	25.4	17.7
10	24.6	8.8	26.0	18.2
15	24.4	8.7	25.8	18.0
20	25.3	9.0	25.6	17.6

Table 5: Automatic metrics (R=ROUGE) of the FlanT5 + TE (PS) model with varying number of persona sentences in the range $[5 - 20]$.

B Perspective Classification

Table 6 presents the results of perspective classification for our top two performing models, compared to the personalized model with average embeddings (Plepi et al., 2022). We report accuracy and the macro F1-score. These metrics are used to evaluate the performance of the model in classifying the perspective (NTA/YTA) based on the generated comments. The previous work is performing better in all splits, due to the model training explicitly for the classification task. Their average embedding model was the highest performing overall, though their priming method achieved 69.6% accuracy on the situation split. However, our FlanT5 + TE (PS) model has a slightly better F1-score by 0.6% in the situation split, which proved to be the most difficult split in the results reported by Plepi et al. (2022). On the other hand, BART + TE (PS), is performing worse in the author split, with a 21% difference compared to the average embedding method, and 10.6% compared to FlanT5 + TE (PS) model.

C Generated Examples

Table 7, shows generated examples from our best models, FlanT5 + TE (PS) and BART + TE (PS) for the same personas per situation. We pick a variety of situations with different perspectives. This table shows the relevance and fluency of the generated outputs with respect to the situation. In addition, we report in Table 8, several comments per situation generated from our model, for different personas. Here we can notice the diversity of generated comments and perspectives per situation, given different personas.

Model	No Disjoint BLEU-1	Acc	F1	Situations BLEU-1	Acc	F1	Authors BLEU-1	Acc	F1
Avg. Embed (Plepi et al., 2022)	-	**86.1**	**83.3**	-	66.5	56.2	-	**86.0**	**83.2**
BART + TE (PS)	9.9	83.4	79.2	3.1	**68.1**	51.2	8.6	74.0	62.2
FlanT5 + TE (PS)	25.3	77.1	73.5	22.2	65.5	**56.8**	23.1	76.2	72.8

Table 6: Perspectivist classification for our best two models and the best model from previous work, the averaged embeddings (Avg. Embed).

Situation	FlanT5 + TE (PS)	BART + TE (PS)
AITA for clapping back on a woman who turned out to be deranged after she made a huge scene on me buying meat products in a grocery store?	NTA - she's gonna learn a valuable lesson from it. It doesn't help her to be a jerk.	NTA. It's not your fault she didn't get help. If she wanted a confrontation she should have done it herself.
AITA for being cold to my ex-fiancé and demanding a paternity test after her infidelity?	NTA. It's an unfortunate situation and the kid is going to end up yours, and you have every right to go out of your way to make it right.	NTA. You're not responsible for her life, and you don't owe her anything. Good on you for sticking up for yourself.
AITA for expecting my wife to do more of the housework because I bring in more income?	YTA There are times where you have to justify your sex life, or try to "compromise" with your partners. Your wife (the A) is willing to do the bulk of the housework and is willing (the B) to work hard	YTA. You are the asshole for expecting your wife to work more hours than you are for asking her to do more work.
AITA for choosing my girlfriend over my parents?	NTA. Your parents are toxic, they will only ruin your life the next time. And it is the least that you could do.	NTA. Your parents are the assholes.
AITA for not wanting to cover/swap shifts with coworkers who are 'sick' with period pain?	NTA They need to see a doctor and be able to manage their own period pain. Just tell them to go get a different doctor and take care of yourself first.	YTA. You don't have to cover someone else's shift just because they're in pain. It's not like you're a bad person for not taking care of your coworkers.
AITA for being angry at my sister for not attending our fathers funeral because of his criminal history?	YTA. She is a convicted murderer, not you. She is angry at him.	YTA. It's your fault not hers.

Table 7: We show the generated outputs from two of our top models, for different situations.

Situation	FlanT5 + TE (PS)
AITA for dating a girl because of her sexual history?	NTA It's reasonable for you to not be interested in a particular type of person YTA. Regardless of your feelings and intentions, you didn't date her because of her sex life. YTA, thats a huge asshole move. If you wanted to break up with her, you should have had a talk with her.
AITA for saying no to celebrating my nephews birthday at my wedding?	NTA, that's your wedding to do as you please. YTA. She's already told you she wouldn't have a big bday celebration
AITA for not letting my partner see my search history?	NTA, in all honesty and trust you are 100% secure, in the world and trust should never be abused NTA -your girlfriend sounds weirdly controlling and untrusting YTA. my husband does this too, and he is probably hiding stuff. I guess you too.

Table 8: We show the generated outputs from our top model, for different situations and different personas.