Perspective Taking through Generating Responses to Conflict Situations

Joan Plepi and **Charles Welch** and **Lucie Flek**
Conversational AI and Social Analytics (CAISA) Lab
University of Bonn, Germany
http://caisa-lab.github.io
{plepi,cfwelch,flek}@bit.uni-bonn.de

Abstract

Although language model performance across diverse tasks continues to improve, these models still struggle to understand and explain the beliefs of other people. This skill requires perspective-taking, the process of conceptualizing the point of view of another person. Perspective taking becomes challenging when the text reflects more personal and potentially more controversial beliefs. We explore this task through natural language generation of responses to conflict situations. We evaluate novel modifications to recent architectures for conditioning generation on an individual's comments and self-disclosure statements. Our work extends the Social-Chem-101 corpus, using 95k judgements written by 6k authors from English Reddit data, for each of whom we obtained 20-500 self-disclosure statements. Our evaluation methodology borrows ideas from both personalized generation and theory of mind literature. Our proposed perspective-taking models outperform recent work, especially the twin encoder model conditioned on self-disclosures with high similarity to the conflict situation.

1 Introduction

Despite the steadily increasing performance language models achieve on a wide variety of tasks, they continue to struggle with theory of mind, or the ability to understand the mental state of others (Carlson et al., 2013; Langley et al., 2022). A range of theory of mind tasks have been studied in developmental psychology, including diverse desire, diverse belief, knowledge-access, false belief, and hidden emotion (Wellman and Liu, 2004). Results consistently show that the hidden emotion, or understanding the difference between felt and displayed emotions, is the most difficult of these tasks (Wellman, 2018). Language assists in the development of theory of mind, as it facilitates the exploration of mental states (Astington and Baird, 2005).
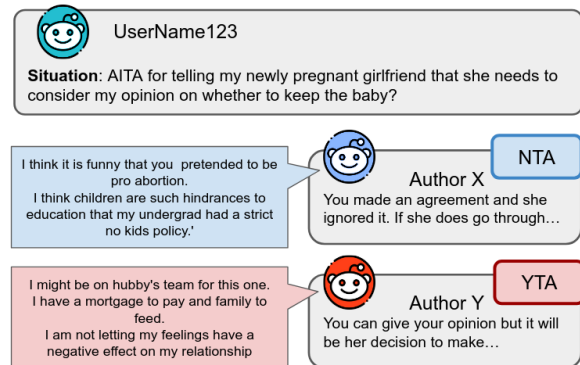

Figure 1: Example of a post in AITA subreddit. The example includes a situation title and two comments with different perspectives regarding the situation, plus self-disclosure sentences for the respective users. Verdicts: NTA = not the asshole, YTA = you are the asshole.

This ability is central to much of human interaction and could provide many benefits for language models as well, as being able to foresee the reactions of others allows us to better decide which action to take next. This could help language models generate responses that are safer (Bommasani et al., 2021), in particular for healthcare applications (Ma et al., 2023), or more personalized, e.g. to sound more empathetic (Lahnala et al., 2022), or provide targeted explanations (Li et al., 2021). In fact, 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; Sorensen et al., 2024). 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, 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 natural language generation (NLG).

We construct a corpus to study perspective taking through generating responses to conflict situations. An example from our 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. On the right, there are two responses from other users with different judgments of the situation (reasoning and verdict NTA/YTA). On the left, we see self-descriptive statements of each user. Author Y appears to be more family-oriented than Author X which may impact their judgement of the situation. Our first research question is, **RQ1**: *How should we evaluate perspective taking through the lens of NLG?* We develop a novel evaluation of by asking humans to rank the human response, model output, and a distractor human response, combining approaches from persona consistency (Madotto et al., 2019) and theory of mind work (Sap et al., 2019).

Next, we ask **RQ2:** *Do tailored, user-contextualized architectures outperform large language models (LLMs) on this task?* We design two transformer architectures to embed personal context and find that our twin encoder approach outperforms LLMs. Lastly, we ask **RQ3:** *What user information is most useful to model perspective taking?* Experiments with varied user context showed that self-disclosure statements semantically similar to the conflict situation were most useful.

Furthermore, by training a generative model to produce perspectives, we outperform recent supervised data perspectivist models also on verdict prediction. Our corpus contains 95K judgements of conflict situations written by 6K authors filtered from Social-Chem-101 and is extended with 20-500 self-disclosure statements per author. We will release our corpus, code, and human evaluations.

2 Related Work

Perspective-taking is part of the process of understanding others mental states and theory of mind is fundamentally based on the understanding that individuals have different beliefs and emotions. Benchmarks for theory of mind in artificial intelligence sometimes take the form of multiple choice questions about false belief (Sap et al., 2019). These are context specific, as a situation and knowledge of involved parties is described, but not unique to individual people. When we look at emotions, the most difficult of theory of mind tasks for children (Wellman, 2018), different people may feel

differently because of their unique perspective and experiences. As such, our work relates to work on personalized understanding and generation.

Recent models, such as RoBERTa or GPT-3, struggle with solving simple perspective-taking toy tasks (Shapira et al., 2023; Trott et al., 2023; Nematzadeh et al., 2018), including the false belief test (Astington and Olson, 1995). Currently, the best LLMs (such as GPT-4 or Llama2-70B) are to a limited extent capable of perpective-taking-based reasoning and chain-of-thought explanations in social commonsense scenarios (Gandhi et al., 2023; Jin et al., 2022; Sap et al., 2022). However, these attempts has been heavily criticized by psychologists as over amplified by the AI community while very basic in design, likely exhibiting shortcut learning (Shapira et al., 2023; Ullman, 2023). Children's theory of mind ability appears associated with "exposure to rich discourse about thoughts, feelings, and intentions" (Hughes and Devine, 2015).

Personalized Generation The first step towards successful modeling of perspective taking is being able to mimic the communication of a given person accurately. Researchers largely agree on the need of including personal, social, cultural, and situational factors into language interpretation and generation, in order to avoid harmful consequences and to steer human-centered AI aligment (Hovy and Yang, 2021; Soni et al., 2022a; Dudy et al., 2021). However the approaches to what and how to model vary dramatically, from an authorial style to chat persona's factual trivia.

Architectural predecessors of personalized LLMs are persona-based dialog systems, attempting to generate a response given an input utterance and additional personal trivia information. Li et al. (2016b) introduce a speaker model and a speaker-addressee model. Madotto et al. (2019) use only a few dialog samples to generate personalized responses, by casting personalized dialog learning as a meta-learning problem. Other works have modified sequence-to-sequence frameworks to infuse persona information in the decoder (Zheng et al., 2019, 2020). Song et al. (2019) introduce Persona-CVAE and Ma et al. (2021) combine a history encoder, personalized post encoder, user history memory, and personalized decoder to fuse the user profile into the response. Recently, Soni et al. (2022b) introduced HaRT, a large-scale transformer model containing a user-state attention layer. Many personalized generation models operate on the Per-

sonaChat dataset (Zhang et al., 2018), where two crowdworkers converse with each other, attempting to represent a persona described by five short sentences, resulting in artificial conversations. Dialogs directly using words from persona sentences were later adjusted (Dinan et al., 2020), however, the data still includes unrealistic personas ("to stay fit, I chase cheetah in the zoo") and unrealistic responses that enforce the usage of the trivia ("I am reading a book."-"Ok. I am a dentist."). Other data include personalized recipes (Majumder et al., 2019) or movie dialogs (Vincent et al., 2023).

3 Dataset of Social Situations

We used the dataset of Welch et al. (2022b) as the foundation of our work, as it addresses the emotional and social aspects of perspective taking, while requiring knowledge of personal beliefs. 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 Extracting Self-Disclosures Statements

We expand the dataset by retrieving the comment histories for each user in the dataset. To extract the self-disclosure sentences for the users, we adapt the approach described in Mazaré et al. (2018). Initially, we split each comment into a sentence

---

[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/`).

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 self-disclosure 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 self-disclosure statements for each user. We filtered our dataset to include only users who have more than 20 and fewer than 500 statements. Our final dataset contains 20K posts and 95K verdicts written by 6K different authors, which we will release upon publication.

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 self-disclosures. 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) auxiliary information. The auxiliary information can be self-disclosure statements (S) or comments (C), and allows us to address **RQ3**: What information is most useful for perspective taking?

In order to tackle this research question, we formalize the following task. For a given situation post s written from a random author, 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 personal context for author a. We describe two different methods to extract a set of k disclosure statements for each user in the dataset.

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

Most relevant sampling We compute embeddings using SBERT (Reimers and Gurevych, 2019), for all extracted self-disclosures and situation titles in our dataset. We compute the cosine similarity
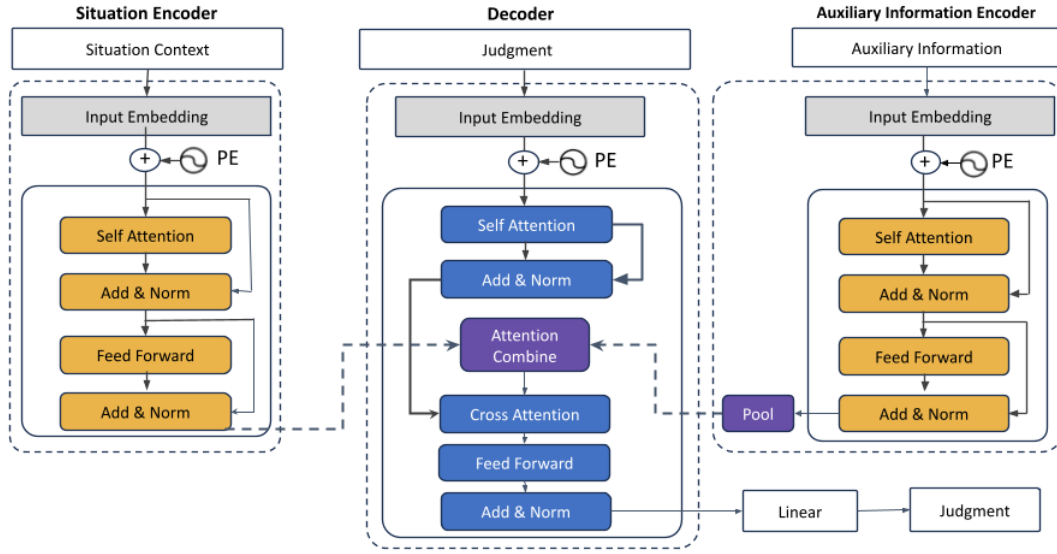
Figure 2: Twin encoder model, with an extra encoder to model the auxiliary user information.

between an author's statements and the situations that they have commented on and select the top k most similar statements for each situation. We aggregate the top k across situations for each author and rank the statements by their frequency, again keeping the top k.

5 Methodology

After formally describing the base transformer (5.1), we outline the modifications implemented to the encoder-decoder model in order to incorporate auxiliary information. We compare our twin encoder (5.2) and style decoder (5.3) models to large language models to answer **RQ2**: Do tailored architectures outperform LLMs?

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:

$$o = decoder(h; \theta^{(dec)})$$
$$\hat{y} = softmax(\mathbf{W_o}^\top o) \tag{1}$$

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, the twin encoder. As we described in §4, we 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 personal context for user a 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 auxiliary sentences, and n_p is the maximum token length in the 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 self-disclosure 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. Finally, $\bar{z}$ is forwarded to the decoder layer, to perform cross attention with the decoder state.
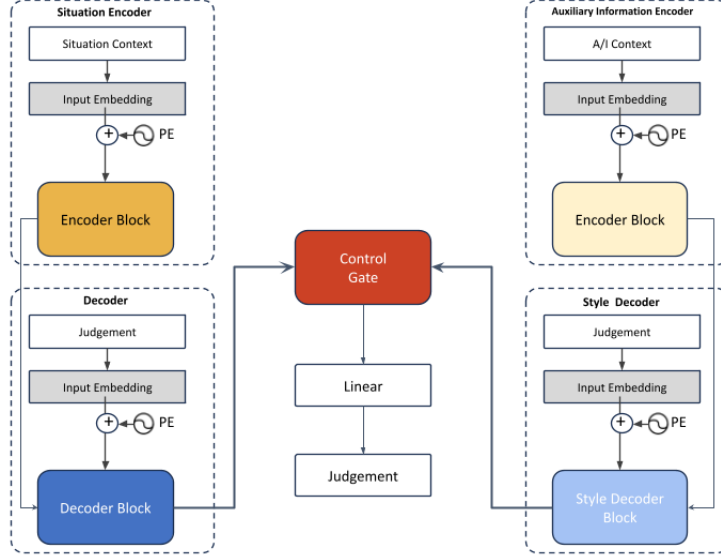
Figure 3: Style decoder model, with a decoder that focuses on persona style, and a control gate that controls the amount of information used from both decoders.

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 auxiliary information. 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:

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

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).

5.4 Large Language Models

In addition to fine-tuning, we explore prompting large transformer models of the Llama2 family (containing 7 to 70 billions of parameters) by utilizing auxiliary information in the prompts. In the first setup, we adjust the prompt in order to include up to 10 examples of auxiliary information, either self-disclosure sentences or comments. Afterward, we utilize pairs of past situation titles and comments of an author to construct the prompts for the models. Moreover, we also report the results of fine-tuning the Llama2-7B model on our task.

6 Experiments

In our experiments, we utilize two base models, that follow an encoder-decoder architecture. We also compare to three recent methods for incorporating personalized information:

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., 2016c; Welch et al., 2022a). However, one drawback of this method is that it cannot generalize to unseen users during test time.

PAA Model. 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 self-disclosure sentences as an auxiliary input.

6.1 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 auxiliary information separately leads to better performance, while for the style decoder, it is concatenated to create a long context. For prompting, we utilize three versions of Llama2 models (Touvron et al., 2023), 7, 13, and 70 billion parameters. An example of the prompt for Llama2 models is given in Table 12 in the Appendix. 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. In order to fine-tune Llama2-7B model, we employ LoRA (Hu et al., 2021) with r=64. For this experiment, we use 4 NVIDIA A100 80GB GPU.

6.2 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. 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. Our development of the human evaluation answers **RQ1**: How should we evaluate perspective taking through the lens of NLG?

In the first part, we focus on matching self-disclosures 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 set of self-disclosure statements when presented with the same set provided to the model. We found that in almost every case across 50 examples and two annotators, the answer was yes. This evaluation is insufficient for our task where it is unlikely for self-disclosures to be directly stated or even rephrased in someone's comments; there is a less direct connection between beliefs reflected in written verdicts and the personal background reflected in self-disclosures.

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

Model	BLEU-2 ↑	Dist-2 ↑	DistS-2 ↑
PAA (Huang et al., 2023)	5.1	0.06	0.53
BART + Priming	1.9	0.14	0.61
BART + User Id	1.7	0.15	0.63
BART + TE (S)	4.2	0.17	0.57
BART + TE (C)	2.4	0.14	0.62
BART + SD (S)	2.0	0.15	0.55
BART + SD (C)	2.4	0.16	0.63
FlanT5 + Priming	4.2	0.1	0.75
FlanT5 + User Id	5.7	0.14	0.77
FlanT5 + TE (S)	**9.0**	**0.387**	**0.92**
FlanT5 + TE (C)	2.9	0.25	0.73
FlanT5 + SD (S)	5.1	0.29	0.8
FlanT5 + SD (C)	5.9	0.29	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-2 score in the range of 0-100 and diversity metrics Distinct 2 grams (Dist-2), and Distinct 2 grams across situations (DistS-2) in the range 0-1. (S) means the model uses self-disclosure sentences as additional information, (C) past comments.

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 self-disclosure statements." Ranking with both correct and incorrect human responses allows us to more clearly understand model performance. This controls for the fact that multiple perspectives can be plausible for a given situation. The model must connect the perspective and the personal background. It is more difficult for models to be ranked over the ground truth than it is to outperform other human 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 about the given situation summary.

7 Results and Analysis

Similar self-disclosure statements were most useful for perspective taking. In Table 1, we report the automatic results for all combinations of architectures from our models. Furthermore, BART-

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
25	24.1	8.3	25.0	17.5
30	22.9	7.7	24.7	17.4

Table 2: Automatic metrics (R=ROUGE) of the FlanT5 + TE (PS) model with varying number of self-disclosure sentences in the range $[5 - 30]$.

based models were the most sensitive with respect to the retrieval method used to extract the set of auxiliary sentences. When random self-disclosure sentences or 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.

In Table 2, we report the results for FlanT5 + TE (PS), with different numbers of self-disclosure sentences as context. Our experiments are run with $\{5, 10, 15, 20, 25, 30\}$. We notice that the best-performing model uses 20 self-disclosure 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 self-disclosure sentences. Incorporating more information is helpful up to a point where the context gets longer and is including less similar statements.

The twin encoder architecture performed best for both models. The key difference between the two architectures is that information about the situation and the auxiliary context is combined. In the twin encoder, 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 general, the FlanT5 variations proved to perform better, which may be attributed to the size difference of the base models (250M vs 140M). 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.[2] Among prim-

---

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

Model	BLEU-2	Dist-2	DistS-2
Llama2-7B FT (S)	**<u>6.5</u>**	0.16	**0.93**
Llama2-7B (S)	5.4	0.1	0.63
Llama2-13B (S)	4.8	0.15	0.82
Llama2-70B (S)	5.8	0.14	0.49
Llama2-7B (C)	4.6	0.053	0.8
Llama2-13B (C)	6.4	0.22	0.9
Llama2-70B (C)	**<u>6.5</u>**	**0.27**	0.81
Llama2-7B (P)	3.5	0.02	0.77
Llama2-13B (P)	4.5	0.12	0.88
Llama2-70B (P)	6.4	0.25	0.86

Table 3: Automatic metrics for different Llama2 models prompted with: 1) self-disclosure (S), 2) comments (C), 3) pairs of past situation/comments, and fine-tuned (FT) version of Llama2-7B model.

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

Table 4: 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.

ing and user ID, that do not require any architecture changes, priming was better. However, in the case of FlanT5 + priming, it generated excessively long responses resulting in nonsense judgments.

LLMs performed worse than tailored architectures. Table 3 shows the results of Llama2 variants for different prompts. Large versions of Llama2 13B and 70B models perform best when the prompt contains past examples of user comments (C). On the other hand, the performance drops for both 7B and 13B models, when utilizing pairs of past situation/comments of the author. This drop in performance may be attributed to the expanded context size resulting from the incorporation of past situations. Additionally, we also report the results for the fine-tuned Llama2-7B model with self-

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

Table 5: Human evaluation results for our top two models BART and FlanT5 fine-tuned with Twin Encoder (TE) with self-disclosure sentences (S), and FlanT5 + Style Decoder (SD), with comments.

disclosure sentences. Performance of the model is improved compared to in-context learning with prompting, reaching the performance of Llama2-70B model. Nevertheless, all the large models performed worse than our top model, despite having almost 100 times more parameters.

Human evaluation confirms superiority of our model, showing gap between human performance. In Table 4 we show the results for the first part of the survey, which is related more to alignment between the generated response and the self-disclosure sentences 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 self-disclosure sentences of the users (see Appendix C for examples). The generated is ranked over the correct response 42% of the time, showing room for improvement, as we would expect this to be close to 50% if they were indistinguishable. 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 5. 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.

Generated perspectives are competitive with perspectivist classification. As a follow-up analysis, we evaluated our model on the perspective classification task from Plepi et al. (2022) by extracting the labels (NTA/YTA) from the generated comments. We use their splits; the verdict split (default for our generative experiments), situation and author splits, which have disjoint sets of situations and authors respectively, across train, validation, and test.

Table 6 presents the results of perspective classi-

	Verdict			Situations			Authors		
Model	BLEU-1	Acc	F1	BLEU-2	Acc	F1	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)	4.2	83.4	79.2	3.1	**68.1**	51.2	8.6	74.0	62.2
FlanT5 + TE (PS)	9.0	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).

fication for our top two performing models, compared to the personalized model with average embeddings (Plepi et al., 2022). The previous work performs better on author and verdict splits. It has the advantage of 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, which proved to be the most difficult split. Our FlanT5 + TE (PS) model has a slightly better F1-score for the situation split

8   Conclusions

As we make progress on improving perspective taking ability, we will need to have models that take additional contextual information into account, especially personal contextual factors. We examined this through the lens of generating perspectives on conflict situations. First, we studied which information is most useful to perspective taking, finding that self-disclosure statements that were similar to the situation were most useful. Second, we compared tailored architectures to LLMs, including two novel methods, finding that our twin encoder architecture outperformed recent work, FlanT5 and Llama2 models. Lastly, we found that previous consistency evaluation metrics were inadequate and proposed a human ranking evaluation that includes similar human responses. Additionally, we found that our generation model performed competitively with previous work on perspective classification. We extended an existing corpus with 20-500 self-disclosure sentences per author and will release our code and corpus upon publication.

Limitations

In this work, we utilize self-disclosure sentences extracted from Reddit in order to improve personalized judgment generation in social media. However, there are a lot of self-disclosure sentences available per user. Even though we attempted to sample the most relevant 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 extracted statements. Moreover, in this work, we train and modify only base models, instead of large ones, due to computation resources.

Performing human evaluation using the self-disclosure 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 sentences. 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 self-disclosure 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 should 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 models.

References

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.

Janet Wilde Astington and Jodie A Baird. 2005. *Why language matters for theory of mind*. Oxford University Press.

Janet Wilde Astington and David R Olson. 1995. The cognitive revolution in children's understanding of mind. *Human development*, 38(4-5):179–189.

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.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Stephanie M Carlson, Melissa A Koenig, and Madeline B Harms. 2013. Theory of mind. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(4):391–402.

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*.

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.

Kanishk Gandhi, J-Philipp Fränken, and Tobias Gerstenberg Noah D Goodman. 2023. Understanding social reasoning in language models with language models. *arXiv preprint arXiv:2306.15448*.

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

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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.

Claire Hughes and Rory T. Devine. 2015. A social perspective on theory of mind. *Handbook of child psychology and developmental science, socioemotional processes*, 3:564.

Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. 2022. When to make exceptions: Exploring language models as accounts of human moral judgment. *Advances in neural information processing systems*, 35:28458–28473.

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.

Allison Lahnala, Charles Welch, David Jurgens, and Lucie Flek. 2022. A critical reflection and forward perspective on empathy and natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Christelle Langley, Bogdan Ionut Cirstea, Fabio Cuzzolin, and Barbara J Sahakian. 2022. Theory of mind and preference learning at the interface of cognitive science, neuroscience, and ai: A review. *Frontiers in Artificial Intelligence*, 5:62.

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.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016c. 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.

Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized transformer for explainable recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4947–4957, Online. 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.

Zilin Ma, Yiyang Mei, and Zhaoyuan Su. 2023. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In *AMIA Annual Symposium Proceedings*, volume 2023, page 1105. American Medical Informatics Association.

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.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400.

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.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*.

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. 2022a. Human language modeling. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 622–636, Dublin, Ireland. Association for Computational Linguistics.

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

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.

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*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Sean Trott, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. 2023. Do large language models know what humans know? *Cognitive Science*, 47(7):e13309.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

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*.

Henry M Wellman. 2018. Theory of mind: The state of the art. *European Journal of Developmental Psychology*, 15(6):728–755.

Henry M Wellman and David Liu. 2004. Scaling of theory-of-mind tasks. *Child development*, 75(2):523–541.

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.

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.

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   Full Results

In Table 7, we present the extended metrics for our models, including ROUGE scores, together with distinct metrics for 1-grams and BLEU-1. It's observed that the pattern of performance across nearly all other metrics aligns with our earlier findings, highlighting that the FlanT5 + TE (PS) model demonstrates superior performance.

Moreover, in Table 8, we report the similarity of the generated text with the auxiliary sentences provided. We provide the maximum, mean, and minimum similarity between the verdict and the sentences in the extracted set. In addition, we provide as a reference, the similarity between the original judgments and the auxiliary information. We notice that our FlanT5 + TE (S) model has the best results and the closest to the gold reference.

B   PersonaChat Discussion

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. Our dataset is instead constructed from the profiles of real people who wrote both the judgements of social situations and their persona sentences.

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 (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 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, the similarity between the persona sentences and responses is high.

C   Generated Examples

Table 9, shows generated examples from our best models, FlanT5 + TE (S) and BART + TE (S) 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 10, 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.

Furthermore, we note in Table 11, samples of self-disclosure sentences from a user in our dataset, together with the generated texts from our FlanT5 + TE (S) model and the original comment. One can

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 (S)	9.9	4.2	25.4	**19.7**	0.033	0.17	0.5	0.57
BART + TE (C)	5.0	2.4	18.8	15.6	0.029	0.14	0.52	0.62
BART + SD (S)	4.2	2.0	19.1	15.8	0.03	0.15	0.41	0.55
BART + SD (C)	5.8	2.4	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 (S)	**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 (S)	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 7: 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 Distinct n-grams (Dist-n), and Distinct n-grams across situations (DistS-n) in the range 0-1. (S) means the model uses self-disclosure sentences as additional information, (C) past comments. The auxiliary set of information is extracted using the most similar method.

Model	Max ↑	Mean ↑	Min ↑
Gold Reference	0.31	0.1	-0.03
PAA (Huang et al., 2023)	0.26	0.08	-0.05
BART + Priming	0.26	0.08	-0.05
BART + User Id	0.25	0.07	-0.05
BART + TE (S)	0.27	0.08	-0.05
BART + TE (C)	0.27	0.08	-0.05
BART + SD (S)	0.25	0.07	-0.05
BART + SD (C)	0.26	0.08	-0.05
FlanT5 + Priming	0.25	0.08	-0.05
FlanT5 + User Id	0.25	0.08	-0.05
FlanT5 + TE (S)	**0.29**	**0.1**	**-0.04**
FlanT5 + TE (C)	0.25	0.08	-0.05
FlanT5 + SD (S)	0.26	0.08	-0.05
FlanT5 + SD (C)	0.26	0.08	-0.05

Table 8: We provide the maximum (Max), mean and minimum (Min) similarity between the verdict and the auxiliary sentences in the extracted set.

infer from the sample that the person giving the judgment had kids while in an intimate relationship and is no longer in that relationship. They give a judgment in similar situations related to these experiences, the first one involving an ex, and the other one involving marriage. The first situation, is a controversial one, with a ratio of original NTA/YTA verdicts equal to 0.47. We notice that in both cases, our model correctly predicts the verdict, together with the rationale behind it.

Situation	FlanT5 + TE (S)	BART + TE (S)
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 9: We show the generated outputs from two of our top models, for different situations.

Situation	FlanT5 + TE (S)
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 10: We show the generated outputs from our top model, for different situations and different personas.

Self-Disclosure Sentences:
1. I think once my kids get busy with activities and sports and things it will be like oh,
dad used to date her and she had kids... But given the distance I don't see them maintaining a relationship.
2. I do have a college fund that is funded by my side of the family for them - my ex doesn't contribute to that.
3. I like the idea of simply thanking her and being honest that you can't reciprocate right now, but appreciate the thought.
4. Just to get it off my plate and to also set the boundary that if it's not emergent,
it's not something I will respond to immediately.
5. I always tell my kids that it's ok to like something and not like other things,
but it's not ok to be dismissive to others of what they enjoy.
6. I explained to my kids that all parents (whether it's a friend's mom or their dad) have different rules for their homes.

Situation A:
AITA for going into my ex-mother-in-law's to chit chat after dropping my kids off?
Generated:
NTA. Your ex and MIL are being very childish. Your ex didn't want you and his mom to talk so much.
There was no reason for her not to be friendly with you.
Gold:
NTA. Here's why I think this is true, in a divorce everyone is hurt, everyone has an opinion about
what relationships need to be maintained and cut off. Your ex-MIL still wants to be friendly with you,
that's her position. She invited you in. You still enjoy talking to your ex-MIL so you accepted reluctantly.
Your ex does not like that you and his mom are friendly, which is fair on his part too, he has another relationship.

Situation B:
AITA for not wanting to attend my friend's wedding since I am not a bridesmaid?
Generated:
NTA. I understand the bridesmaids feeling you wanted to be there for her, but you can't choose what her party is.
That's just a problem for her.
Gold:
NTA. I don't think it's wrong to avoid the wedding considering you state it's a financial burden
that you'd rather avoid if you aren't in the wedding. You thought your friendship was in a different place than she did.
She's not picked you, not an asshole for being upset - you are entitled to your feelings.

Table 11: Sampled self-disclosure sentences from a user in our dataset, together with the generated texts from our FlanT5 + TE (S) model and the original comment.

[INST] <<SYS>>
I will provide self-disclosure sentences written from a user. Generate a perspective for the given situation.
Perspective should be aligned with the user inferred personality. The perspective should start with the verdict YTA or NTA.
YTA means You're the asshole and NTA means Not the asshole.<</SYS>>
Sampled self-disclosure sentences for user:
I don't think that talking is wrong.
I just tried the game.
I don't think that your wife will ever be completely faithful or honest with you.
I think it was lovely for you to not only offer beef stew, but also to offer to make vegetable stew for the non-beef-eating daughter.
I think you might be overreacting, because it's normal for to follow someone on multiple social media accounts.
I would, in your shoes, tell my MIL that you will cut contact with her if she doesn't respect your DD's boundaries.
I want to go in a date with my husband.
Yes, please help me maintain my lie so that I can date this guy.
I don't blame you for not going after them for the damages.
Yes, I have been married to my amazing husband for 24 years.
Situation: AITA for telling a girl that I wasn't looking for a relationship?
Users Perspective: [/INST]

NTA as long as you politely told her that you're just not interested in a relationship right now.
And by "politely" I mean you didn't put her down in the last text.

Table 12: An example from our prompt with self-disclosure sentences for Llama2 models.