Harmonizing Code-mixed Conversations: Personality-assisted Code-mixed Response Generation in Dialogues

Shivani Kumar
IIIT Delhi
shivaniku@iiitd.ac.in

Tanmoy Chakraborty
IIT Delhi
chak.tanmoy.iit@gmail.com

Abstract

Code-mixing, the blending of multiple languages within a single conversation, introduces a distinctive challenge, particularly in the context of response generation. Capturing the intricacies of code-mixing proves to be a formidable task, given the wide-ranging variations influenced by individual speaking styles and cultural backgrounds. In this study, we explore response generation within code-mixed conversations. We introduce a novel approach centered on harnessing the Big Five personality traits acquired in an unsupervised manner from the conversations to bolster the performance of response generation. These inferred personality attributes are seamlessly woven into the fabric of the dialogue context, using a novel fusion mechanism, PA3. It uses an effective two-step attention formulation to fuse the dialogue and personality information. This fusion not only enhances the contextual relevance of generated responses but also elevates the overall performance of the model. Our experimental results, grounded in a dataset comprising of multi-party Hindi-English code-mix conversations, highlight the substantial advantages offered by personality-infused models over their conventional counterparts. This is evident in the increase observed in ROUGE and BLUE scores for the response generation task when the identified personality is seamlessly integrated into the dialogue context. Qualitative assessment for personality identification and response generation aligns well with our quantitative results.

1 Introduction

Conversations[1] serve as the primary medium for exchanging ideas and cultivating acquaintance among individuals (Turnbull, 2003). Remarkably, many people exhibit fluency in multiple languages, seamlessly blending these linguistic resources in their daily communications (Tay, 1989; Tarihoran and Sumirat, 2022). This phenomenon, characterized by fusing distinct languages to convey meaning, is referred to as *code-mixing*. While code-mixing prevails as a widespread linguistic phenomenon (Kasper and Wagner, 2014), it has not garnered significant attention within the mainstream NLP community, where monolingual text processing has been the predominant focus. Of late, there is a growing recognition of the critical importance of comprehending code-mixed conversations resulting in an increased number of studies investigating diverse aspects of code-mixing in conversations (Banerjee et al., 2018; Agarwal et al., 2021; Singh et al., 2022; Dowlagar and Mamidi, 2023), such as the identification of humor (Khandelwal et al., 2018; Bedi et al., 2021; Bukhari et al., 2023), emotional expression (Ameer et al., 2022; Kumar et al., 2023b), and sarcasm (Bedi et al., 2021; Kumar et al., 2022). However, the realm of response generation within code-mixed dialogues remains an underexplored frontier (Singh et al., 2022). To this end, we propose tackling the response generation challenge for code-mixed conversations.

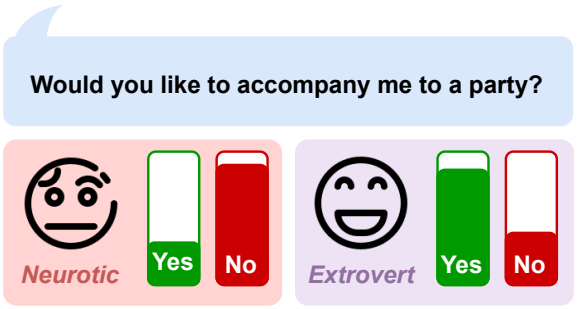


Figure 1: Influence of personality on dialogue responses – a *neurotic* speaker might respond negatively to the posed question, whereas an *extrovert* would likely provide a positive reply.

It is crucial to note that while response generation is a vital avenue to explore, it diverges significantly from conventional natural language under-

[1]We use 'conversations', 'dialogues', and 'discourse' interchangeably.

Findings of the Association for Computational Linguistics: EACL 2024, pages 639–653

standing tasks since a uniform, 'one-size-fits-all' model proves inherently inadequate in this context (Chen et al., 2020a). Every individual possesses a unique set of preferences and life experiences, which collectively mould their distinct personalities, subsequently exerting a profound influence on their responses to identical queries (Zhang et al., 2018a). Figure 1 illustrates this point. As evident, the response to a seemingly straightforward question, such as *"Would you like to accompany me to a party?"*, can differ based on the listener's prominent personality traits. Interlocutor A, characterized as an *neurotic*, responds distinctively compared to Interlocutor B, who leans more towards being *extrovert*. Appendix A.1 presents the definition of personality traits, along with examples.

Personality traits, by their very nature, span a vast spectrum and thus possess the potential for infinite possibilities (Alam and Riccardi, 2014). Numerous studies have been conducted to quantify these traits (Briggs and Myers, 1995; Butcher and Williams, 2009; Benjamin Jr, 2020), with the Big Five personality traits (Digman, 1990) emerging as the prominent framework in this context. This theory distils human personality into five distinctive dimensions: Openness (OPN), Conscientiousness (CON), Extraversion (EXT), Agreeableness (AGR), and Neuroticism (NEU), in which each dimension encapsulates a pivotal facet of an individual's character. For instance, elevated levels of openness may signify a predisposition towards imagination. Here, we adopt this widely accepted model as the foundation for characterizing a speaker's personality. Our central hypothesis contends that incorporating personality indicators within the response generation process plays a pivotal role in generating contextually appropriate responses to given queries. Given the intricate and non-generalizable nature of manually annotating personality traits, we propose an unsupervised learning approach to acquire these traits, which, in turn, enhances response generation capabilities. In a nutshell, our contributions are four-fold:

1. We explore the task of **code-mixed response generation**.
2. We propose an **unsupervised mechanism to identify speakers' personality traits** and leverage them for better response generation.
3. We propose **a novel method, PA3**[2], which combines the identified traits with dialogue context to generate responses.
4. Our **quantitative and qualitative analyses** show the benefits of including personality traits in code-mixed response generation.

[2]Personality-Aware Axial Attention

2 Related Works

Conversation and Code-mixing. Dialogues represent a well-established domain in NLP, having undergone extensive exploration (Chen et al., 2017; Kumar et al., 2023a). However, the bulk of this research has predominantly revolved around monolingual text, despite the enduring prevalence of code-mixing, a timeworn linguistic phenomenon (Tay, 1989). Consequently, recent years have witnessed a surge in studies dedicated to unravelling the intricacies of code-mixing within dialogues (Ahn et al., 2020). These investigations have honed in on exploring various nuances of code-mixed dialogues, delving into attributes such as intents (Liu et al., 2020c; Firdaus et al., 2023), the presence of hate speech (Modha et al., 2021; Madhu et al., 2023), humor (Khandelwal et al., 2018; Bedi et al., 2021), and sarcasm (Bedi et al., 2021; Kumar et al., 2022). Yet, the landscape for the generative dimension of code-mixing remains relatively uncharted, with limited concerted efforts in this direction.

Response Generation. For dialogue agents, it is of paramount importance to keep the conversation engaging (Gottardi et al., 2022). Consequently, generating apt responses becomes a primary field of research in terms of dialogue analysis. Many studies have been conducted to generate the right responses for monolingual English dialogues (Spring et al., 2019; Fan et al., 2020; Dong et al., 2022). However, response generation in the code-mixed setting remains a comparatively unexplored topic with only a handful of existing studies (Agarwal et al., 2021; Singh et al., 2022). Bawa et al. (2020) illustrated that multilingual speakers prefer chatbots that can code-mix, thus making code-mixed response generation crucial.

Big Five Personality Traits. In pursuit of a deeper understanding of the user's personality, a range of studies have delved into the realm of the Big Five personality (Costa and McCrae, 1992; Costa Jr and McCrae, 2008). Numerous studies endeavored to categorize individuals into one of these personality archetypes based on their salient attributes (Mairesse et al., 2007; Golbeck et al., 2011; Kosinski et al., 2013; Schwartz et al., 2013). A few studies have also attempted to use different

personality theories other than the Big Five personality traits such as MBTI (Briggs and Myers, 1995; Celli and Lepri, 2018).

Personality-assisted Response Generation. The significance of personalization in enhancing the efficacy of dialogue systems is widely acknowledged (Lucas et al., 2009; Joshi et al., 2017; Weston et al., 2018; Dinan et al., 2018; Roller et al., 2020; Chen et al., 2020b). While earlier studies primarily concentrated on the utilization of user profiles to tailor goal-oriented dialogue systems (Lucas et al., 2009; Joshi et al., 2017), recent investigations have shifted their focus towards chit-chat settings (Li et al., 2016; Zhang et al., 2018b; Weston et al., 2018; Roller et al., 2020; Dinan et al., 2018). However, all of these studies deal with monolingual data. Consequently, we explore personality-assisted response generation in a code-mixed setting.

3 Problem Definition

The complete problem definition can be divided into two phases as follows:

Phase 1: Speaker Personality Detection. Given the contextual utterances $(s_1, u_1), (s_2, u_2), \ldots, (s_{n-1}, u_{n-1})$ such that utterance u_i is uttered by speaker s_j, we aim to generate personality p_n for speaker s_n. A classification model selects p_n, such that $p_n \in P$ and $P = \{\text{OPN}, \text{CON}, \text{EXT}, \text{AGR}, \text{NEU}\}$, and maps the selected trait class into a templatic personality defining the speaker (c.f. Table 1). We append this definition with the input and move on to phase 2.

Phase 2: Response generation. Along with the contextual utterances, the input also contains the personality trait for the subsequent speaker, such that the input becomes $\{(s_1, u_1), (s_2, u_2), \ldots, (s_{n-1}, u_{n-1}), p_n\}$. Response generation aims to generate utterance u_n by speaker s_n based on the detected personality p_n.

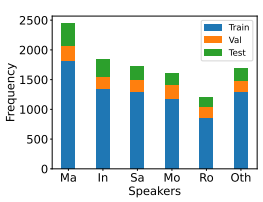
4 Dataset

Datasets for code-mixed conversations are inadequate, especially for multi-turn, multi-party conversations. In this study, we consider the MaSaC dataset (Bedi et al., 2021), containing Hindi-English code-mixed discourse of multi-turn and multi-party nature from an Indian TV series[3]. The dataset was originally curated to perform the task of sarcasm and humour detection since it contains conversations similar to daily discourse among peers.

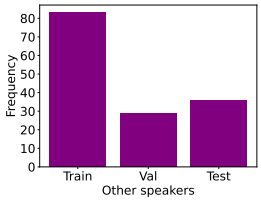
[3] https://www.imdb.com/title/tt1518542/

Set	#Dlgs	#Utts	Avg sp/dlg	Utt len		Vocab len	
				Avg	Max	English	Hindi
Train	8506	8506	3.60	10.82	113	3157	14803
Val	45	1354	4.13	10.12	218		
Test	56	1580	4.32	10.61	84		
Total	8607	11440	12.05	31.55	415		

(a) Statistics of MaSaC.



(b) Speaker distribution in the MaSaC dataset.



(c) Number of speakers other than the five primary speakers.

Figure 2: Dataset description of MaSaC (Abbreviation: Dlgs: Dialogues, Utts: Utterances, sp: speakers, Ma: Maya, In: Indravardhan, Sa: Sahil, Mo: Monisha, Ro: Rosesh, Oth: Others).

Consequently, we extract the conversations from this dataset and construct our response generation instances. We highlight the critical statistics of the dataset in Table 2a. The speaker distribution in Figure 2b and Figure 2c shows that there are five primary speakers in the dataset, each with varying personalities (c.f. Table 2). Thus, aiding response generation with speaker personalities can improve its performance.

5 Proposed Methodology

In this section, we discuss our proposed methodology, with the foremost objective being the effective identification of personality attributes from the dialogue context. To achieve this, we propose an unsupervised technique that leverages response generation performance to improve personality identification. Subsequently, we fuse the personality attributes into the dialogue context to generate responses influenced by individual traits. We propose the incorporation of an intermediary module within the core encoder. This module leverages a straightforward yet effective two-step attention mechanism, facilitating the fusion of personality attributes with the representation of the dialogue. Broadly, we employ context-aware attention (Yang et al., 2019), which is instrumental in infusing personality characteristics into the key and value vectors of the dialogue. Subsequently, we employ Axial attention (Ho et al., 2020) to yield a refined, conclusive representation, which ultimately feeds into the decoder. Figure 4 provides a schematic diagram of our model. In the following subsections, we offer

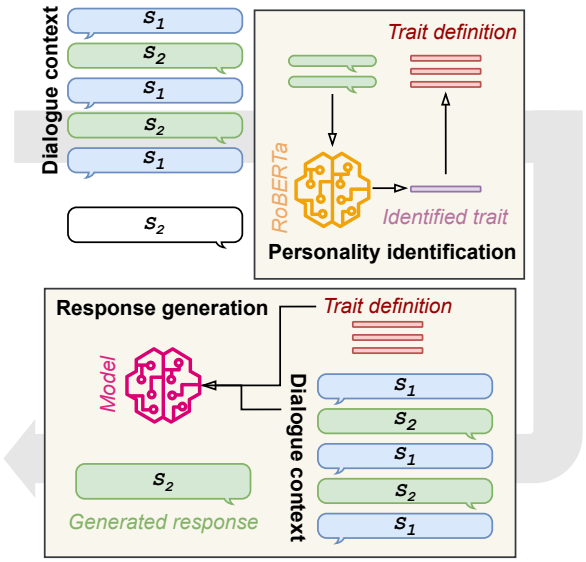


Figure 3: Outline of learning personality traits using the *'pseudo'* task of response generation.

a comprehensive overview of individual modules.

5.1 Personality Identification

In this section, we describe our methodology for discerning the personality traits of each speaker and subsequently mapping them to their corresponding trait definitions. Although multiple theories quantify a speaker's personality traits (Briggs and Myers, 1995; Butcher and Williams, 2009; Benjamin Jr, 2020), existing NLP applications widely use the Big Five Personality theory (Digman, 1990). Consequently, we select this model for our study, encompassing five distinct personality dimensions as shown in Table 1, where one of these dimensions is presumed to be dominant. To find the most suitable personality trait for a speaker in a dialogue, we employ an approach similar to Word2Vec (Mikolov et al., 2013), where a *'pseudo'* task is implemented to facilitate the acquisition of word embeddings. In the context of personality identification, our *'pseudo'* task takes the form of response generation, where we seek to enhance the generated response based on the intermediary step of personality identification. Figure 3 gives an overview of our mechanism for personality identification. We employ RoBERTa base (Liu et al., 2020b) to classify personalities attributed to the target speaker, using the input dialogue as the primary data source. Once the personality is identified, it is subsequently linked to its templatic definition — a descriptive representation of the speaker's character, as outlined in Table 1. This personality definition is presented alongside the input dialogue to an encoder for further steps in the proposed pipeline.

Trait	Templatic Definition
Openness	The speaker has high openness trait. They embrace new ideas, are curious about the world, and are often drawn to creative and unconventional pursuits.
Conscientious	The speaker has conscientiousness trait. They are reliable, organized, and detail-oriented, demonstrating a strong work ethic and a commitment to achieving their goals.
Extraversion	The speaker has extraversion trait. They thrive in social settings, energized by interactions with others, and enjoy being at the center of activities.
Agreebleness	The speaker has agreeableness trait. They prioritize cooperation, are empathetic, and often go out of their way to maintain harmonious relationships and help others.
Neuroticism	The speaker has high neuroticism trait. They have a greater tendency for emotional instability, anxiety, and a propensity to experience negative emotions such as fear, sadness, and anger.

Table 1: Personality traits in the Big Five personality model along with their templatic definitions.

5.2 Personality-Aware Attention (PAA)

With the personality definition and the input dialogue at our disposal, our next step is to seamlessly integrate the personality information with the dialogue information to craft a suitable response. Conventional attention-based fusion mechanisms often facilitate a direct interplay between the input representations, in which one representation functions as the query while the others assume the roles of key and value. However, as each representation captures distinct attributes, straightforward fusion may not preserve the optimal contextual information and could introduce significant noise into the final representations. Consequently, we introduce personality-aware attention (PAA) fusion employing context-aware attention (Yang et al., 2019). Our method entails the initial generation of personality-conditioned key and value vectors, followed by applying axial attention (Ho et al., 2020) to obtain the final fused values. We explain the process in detail below.

For an encoder model, we have the intermediate representation H at a specific layer to compute the query, key, and value vectors denoted as Q, K, and V respectively, in $\mathbb{R}^{n \times d}$ as outlined in Equation 1. $W_Q, W_K, and W_V$ are model parameters each with dimensions of $\mathbb{R}^{d \times d}$. In this context, n signifies the maximum sequence length of the text, while d represents the dimensionality of the dialogue vector.

$$\left[QKV\right] = H\left[W_Q W_K W_V\right] \quad (1)$$

The vector P in $\mathbb{R}^{n \times d_p}$, the encoded personality

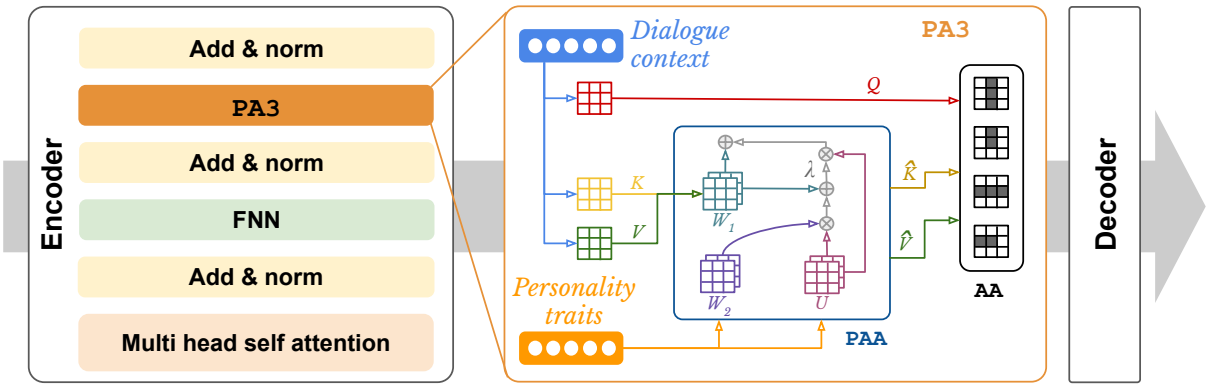


Figure 4: Model architecture to fuse personality values with dialogue context. The PA3 module can be injected into any encoder-decoder architecture, and it takes as inputs the dialogue representation along with the personality trait definition representation. First, context-aware attention is used to learn personality-infused key and value pairs and axial attention is then used to combine query, key, and value vectors into one final representation.

vector is used to create personality-influenced key and value vectors, $\hat{K}$ and $\hat{V}$ respectively, based on the method outlined by Yang et al. (2019). For balancing of information from the personality source and information retention from the dialogue, we train a matrix λ in $\mathbb{R}^{n\times 1}$ (Equation 3). U_k and U_v in $\mathbb{R}^{d_p\times d}$ are matrices that can be learned.

$$\begin{bmatrix}\hat{K}\\ \hat{V}\end{bmatrix} = (1 - \begin{bmatrix}\lambda_k\\ \lambda_v\end{bmatrix}) \begin{bmatrix}K\\ V\end{bmatrix} + \begin{bmatrix}\lambda_k\\ \lambda_v\end{bmatrix} (P \begin{bmatrix}U_k\\ U_v\end{bmatrix}) \quad (2)$$

Rather than setting λ_k and λ_v as hyperparameters, we allow the model to autonomously determine their values through a gating mechanism, as defined in Equation 3. Additionally, the matrices $W_{k_1}, W_{k_2}, W_{v_1}$, and W_{v_2}, each with dimensions $\mathbb{R}^{d\times 1}$, are trained in conjunction with the model.

$$\begin{bmatrix}\lambda_k\\ \lambda_v\end{bmatrix} = \sigma(\begin{bmatrix}K\\ V\end{bmatrix} \begin{bmatrix}W_{k_1}\\ W_{v_1}\end{bmatrix} + P \begin{bmatrix}U_k\\ U_v\end{bmatrix} \begin{bmatrix}W_{k_2}\\ W_{v_2}\end{bmatrix}) \quad (3)$$

Once we obtain the personality-infused key and value vectors, we use the Axial attention mechanism as described below.

5.3 Axial Attention

Axial attention (Ho et al., 2020) finds its primary application in computer vision, where its utility extends to managing multidimensional tensors. The fundamental aim is to approach each axis independently, thereby comprehensively exploring relationships between the various dimensions. The proposed approach preserves the original shape of the multidimensional tensor, performing either masked or unmasked attention along a single axis at any given time. This specific operation, referred to as axial attention and denoted as $\text{Attention}_k(x)$, is responsible for directing attention over axis k within the tensor x. In doing so, it blends information across axis k while maintaining the independence of information along the remaining axes. Implementing axial attention for a given axis k involves a series of steps, such as transposing all axes except k to the batch axis, invoking standard attention as a subroutine, and reverting the transpose operation. Within our network architecture, we leverage two axial attention layers, culminating in the derivation of the ultimate dialogue representation denoted as $\hat{H}$, signifying the personality-infused representation of the dialogue, which is then passed on to the next encoder/decoder layer. For our input two dimensional arrays of $\hat{K}$, $\hat{V}$, and Q:

$$\hat{H} = \text{Attention}_k(\hat{K}, \hat{V}, Q) \quad (4)$$

6 Experiments and Results

Evaluation Metrics. Given the absence of ground-truth labels for evaluating personality detection, we resort to a manual assessment process, meticulously scrutinizing the outputs for the primary speakers to derive meaningful insights into the system's performance in this regard. To assess the response generation proficiency, we employ established evaluation metrics, specifically ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) scores. These metrics are adept at quantifying the syntactic competence of the system in question. Additionally, we incorporate BERTScore (Zhang et al., 2019), which serves to gauge the semantic aptitude of the system, and human evaluation provides a more comprehensive evaluation.

In this section, we present a comprehensive overview of the quantitative and qualitative results achieved by personality identification and response

Sp	GT	OPN	CON	EXT	EXT	NEU
Ma	CON	14%	**54%**	8%	13%	11%
In	AGR	6%	18%	8%	**65%**	3%
Sa	CON	14%	**52%**	4%	16%	14%
Mo	OPN	**58%**	11%	21%	8%	2%
Ro	EXT	16%	14%	**51%**	15%	4%

Table 2: Percentage of times a personality trait is assigned to a speaker. (Abbr - Sp: Speakers, GT: Ground Truth, Ma: Maya, In: Indravardhan, Sa: Sahil, Mo: Monisha, Ro: Rosesh, Oth: Others)

generation. Additionally, we offer a closer look at our ablation results, shedding light on the significance of each submodule within our proposed architectural framework for response generation. Further, human evaluation highlights the pros and cons of the generated responses and personalities.

6.1 Personality Identification

As shown in Figure 3, our initial step predicts the most suitable personality from the Big Five personality traits for the target speaker. To gauge the efficacy of our predicted personalities, we focus on the five primary speakers featured in the MaSaC dataset. Figure 2b shows the distribution of the speakers where it can be observed that the speakers — Maya, Indravardhan, Sahil, Monisha, and Rosesh, are the most frequently occurring speakers. We perform a manual evaluation of the personality predictions. Using information from Wikipedia[4], we procure character descriptions for each of the five prominent speakers (c.f. Appendix A.2) which were given to five expert annotators. The annotators then categorize each speaker within the Big Five personality framework. More information can be found in Appendix A.3. This annotator-driven classification enables the construction of a definitive ground-truth for evaluation encompassing the five speakers, each associated with an assigned personality trait value as shown in Table 2. We compare the obtained ground-truth personalities with the ones predicted by the RoBERTa model, an outcome of the *'pseudo'* task centred around response generation. The ensuing distribution of these predictions is laid out for scrutiny in both Table 2 and Figure 5. We can see that the personalities found most suitable by the human annotators are the ones preferred by the RoBERTa model, too, validating the performance of our system.

[4] https://en.wikipedia.org/wiki/Sarabhai_vs_Sarabhai

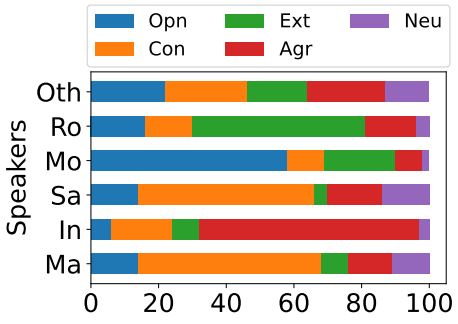


Figure 5: Distribution of the predicted personality traits assigned to different speakers (Abbr - Ma: Maya, In: Indravardhan, Sa: Sahil, Mo: Monisha, Ro: Rosesh, Oth: Others).

6.2 Response Generation

Here, we discuss the effect of adding personality information to the dialogue context quantitatively.

6.2.1 Comparative Systems

To attain the most promising textual representations for discourse, we employ a range of well-established encoder-decoder-based sequence-to-sequence (seq2seq) models. (i) **RNN**: We leverage the RNN seq2seq architecture, implemented through openNMT4[5]. (ii) **Pointer Generator Network (PGN)** (See et al., 2017): In this seq2seq architecture, a fusion of generative and copy mechanisms is harnessed, offering a versatile approach to content generation. (iii) **Transformer** (Vaswani et al., 2017): Responses are generated using the conventional Transformer encoder-decoder model. (iv) **T5** (Raffel et al., 2020): We deploy the base version of the text-to-text-transfer-transformer (T5), which excels in framing multiple NLP tasks as text-to-text challenges, facilitating a unified and efficient approach to tasks such as translation, summarization, and question answering. (v) **BART** (Lewis et al., 2020): We utilize the basic denoising autoencoder model with a bidirectional encoder and a left-to-right auto-regressive decoder. (vi) **mBART** (Liu et al., 2020a): mBART[6], trained on multiple extensive monolingual datasets, shares the same objective and architectural structure as BART.

6.2.2 Quantitative Results

Table 3 presents the scores achieved across the evaluation metrics for the MaSaC dataset. Apparently, the inclusion of personality information elevates the performance of our comparative systems across

[5] https://github.com/OpenNMT/OpenNMT-py

[6] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

	Model	R1	R2	RL	B1	B2	B3	B4	BS
w/o personality	**RNN**	8.17	0.02	8.09	5.11	0.01	0.11	0	54.16
	PGN	7.06	0	7.01	4.31	0	0.08	0	53.12
	Transformers	10.64	0.83	10.35	7.22	0.92	0.13	0.01	58.94
	mBART	11.36	1.23	10.9	7.91	1.01	0.21	0	61.02
	T5	11.87	1.01	11.43	8.41	1.02	0.17	0.02	61.98
	BART	12.94	1.66	12.34	9.66	1.64	0.43	0.07	63.12
w personality	**RNN$_{PA3}$**	9.96 (↑1.79)	0.08 (↑0.06)	10.71 (↑2.62)	6.87 (↑1.76)	1.04 (↑1.03)	0.43 (↑0.32)	0.22 (↑0.22)	56.24 (↑2.08)
	PGN$_{PA3}$	8.45 (↑1.39)	1.11 (↑1.11)	9.41 (↑2.40)	5.95 (↑1.64)	1.03 (↑1.03)	0.37 (↑0.29)	0.21 (↑0.21)	55.87 (↑2.75)
	Transformers$_{PA3}$	12.76 (↑2.12)	1.75 (↑0.92)	12.14 (↑1.79)	8.46 (↑1.24)	2.02 (↑1.10)	0.45 (↑0.32)	0.24 (↑0.23)	61.06 (↑2.12)
	mBART$_{PA3}$	13.43 (↑2.07)	2.36 (↑1.13)	12.15 (↑1.25)	8.89 (↑0.98)	**2.61** (↑1.60)	0.56 (↑0.35)	0.18 (↑0.18)	63.42 (↑2.40)
	T5$_{SC}$	12.02 (↑0.15)	1.51 (↑0.50)	11.98 (↑0.55)	8.52 (↑0.11)	1.51 (↑0.49)	0.39 (↑0.22)	0.11 (↑0.09)	62.05 (↑0.07)
	T5$_{DPA}$	12.04 (↑0.17)	1.56 (↑0.55)	12.01 (↑0.58)	8.58 (↑0.17)	1.58 (↑0.56)	0.41 (↑0.24)	0.14 (↑0.12)	62.35 (↑0.37)
	T5$_{PA3-Axial}$	12.79 (↑0.92)	1.64 (↑0.63)	12.53 (↑1.10)	9.04 (↑0.63)	1.96 (↑0.94)	0.46 (↑0.29)	0.18 (↑0.16)	62.99 (↑1.01)
	T5$_{OT}$	13.48 (↑1.61)	1.97 (↑0.96)	12.89 (↑1.46)	9.21 (↑0.80)	2.23 (↑1.21)	0.52 (↑0.35)	0.21 (↑0.19)	63.14 (↑1.16)
	T5$_{PA3}$	13.61 (↑1.74)	2.03 (↑1.02)	13.92 (↑2.49)	9.78 (↑1.37)	2.62 (↑1.60)	0.51 (↑0.34)	0.26 (↑0.24)	63.87 (↑1.89)
	BART$_{SC}$	13.05 (↑0.11)	1.89 (↑0.23)	12.64 (↑0.30)	9.84 (↑0.18)	1.82 (↑0.18)	0.52 (↑0.09)	0.12 (↑0.05)	63.48 (↑0.36)
	BART$_{DPA}$	13.12 (↑0.18)	1.98 (↑0.32)	12.81 (↑0.47)	9.96 (↑0.30)	1.94 (↑0.30)	0.54 (↑0.11)	0.15 (↑0.08)	63.82 (↑0.70)
	BART$_{PA3-Axial}$	13.97 (↑1.03)	2.21 (↑0.55)	13.05 (↑0.71)	10.16 (↑0.50)	2.07 (↑0.43)	0.61 (↑0.18)	0.18 (↑0.11)	64.34 (↑1.22)
	BART$_{OT}$	14.29 (↑1.35)	2.54 (↑0.88)	13.72 (↑1.38)	10.59 (↑0.93)	2.16 (↑0.52)	0.73 (↑0.30)	0.22 (↑0.15)	65.05 (↑1.93)
	BART$_{PA3}$	**14.67** (↑1.73)	**2.77** (↑1.11)	**14.11** (↑1.77)	**10.92** (↑1.26)	2.51 (↑0.87)	**0.73** (↑0.30)	**0.27** (↑0.20)	**65.93** (↑2.81)

Table 3: Experimental and ablation results for the response generation task with and without fusing personalities. Refer to Appendix A.4 for visualisation (Abbr: R1/2/L: ROUGE-1/2/L, B1/2/3/4: BLEU-1/2/3/4, BS: BERTScore, SC: Simple Concat, DPA: Dot Product Attention, OT: Only Traits, PA3: Personality-Aware Axial Attention).

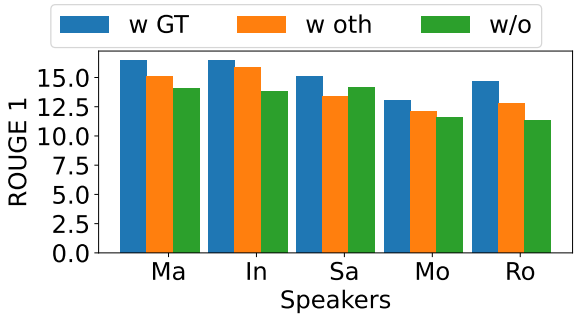


Figure 6: ROUGE-1 scores for the responses generated by the most frequent five speakers in the dataset when the GT personality, other personalities sans GT, and no personalities are used for response generation.

all metrics. Notably, BART outperforms the competition, whether with or without personality information, across majority of the metrics. We observe increased ROUGE-1 scores for all models, typically ranging from $+13\%$ to $+21\%$. BLEU-1 also increases simultaneously from $+12\%$ to $+38\%$. The consistent improvement in BERTScore ($+3\%$ to $+5\%$) also underscores that the fusion of personality information into the dialogue context results in responses marked by enhanced coherence.

6.2.3 Effect of Personality

We monitor ROUGE scores for responses from the top five most frequent speakers, as shown in Figure 2b. Comparing the responses generated by the BART model with ground-truth (GT) personalities (as listed in Table 2), we also assess results without personality fusion. The findings, presented as ROUGE-1 scores in Figure 6, consistently demonstrate improved performance after personality fusion. Notably, except for Sahil, every speaker exhibits enhanced performance when infused with the GT personality within the dialogue context.

6.2.4 Ablation Study

It is essential to recognize that integrating personality information into the dialogue context can be achieved through various techniques, each varying in complexity. In our study, we have delved into several fusion methodologies, encompassing straightforward concatenation, conventional dot-product attention, and personality-aware attention, both with and without the inclusion of Axial attention. We provide results for both BART and T5 since they exhibit comparable capabilities in Table 3. Evidently, the fusion of personality information contributes to better responses. Nevertheless, our findings emphasize that simple concatenation falls short in efficiency, yielding only marginal performance gains. In contrast, introducing attention mechanisms elevates performance, with our proposed approach of personality-aware fusion, coupled with Axial attention, being the most effective strategy. Additionally, we investigate the potential impact of fusing solely the identified personality trait without the intermediary step of mapping it into a trait definition. Our observations underscore the advantages of incorporating the complete trait definition rather than merely the isolated trait string within the response generation pipeline.

Dialogue Context	**Monisha:** Mummy ji, use apni pehli anniversary yaad hai, ye sunkar aap chaunki kyun? *(Mother-in-law, why did you seem surprised when he remembered his first anniversary?)* **Maya:** Kyunki mardon ko, cricket match kab hai, WWF kab dikhane wale hain, inn do baaton ko chhod bahut jyada yaad ni rehta. And that's fine. *(Because men tend to remember when the next cricket match or WWE show is rather than things like anniversaries. And that's fine.)* **Monisha:** Kya fine? Mujhe to har important din yaad rehta hai. *(What do you mean, fine? I remember every important day.)* **Maya:** Haan, kyunki mard aur aurat alag-alag hain. Unke dimag alag tarah se kaam krte hain. *(Yes, because men and women are different. Their minds work in different ways.)*
GT Response	**Monisha:** Mai nhi manti. pichhle saal raat ko barah baje daddy ji ne aapko anniversary wish kiya tha na? Yaad krk? *(I don't believe that. Didn't Daddy wish you a happy anniversary last year at midnight? Remember?)*
BART	**Monisha:** daddy ji ka dimag alag tarah se kaam krte hain. *(Daddy's mind work in different way.)*
$BART_{PA3}$ (OPN)	**Monisha:** Mai nhi manti. daddy ji ko apni pehli anniversary yaad hai. *(I don't believe that. Daddy remembers his first anniversary.)*

Table 4: Responses generated for a sample dialogue from the test set of MaSaC by different model architectures.

Model	Fluency	Coherence	Relevancy	Personality oriented
T5	2.13	2.07	1.64	2.01
BART	2.17	2.03	1.79	2.04
$T5_{PA3}$	3.07	2.84	2.26	3.11
$BART_{PA3}$	**3.14**	**3.09**	**2.98**	**3.23**

Table 5: Results of human evaluation for the response generation task.

6.2.5 Qualitative Analysis

We select a sample dialogue from the test set and present the predicted responses generated by the conventional BART model alongside those generated after the integration of personality factors using PA3. These responses are compared with the ground-truth responses, comprehensively detailed in Table 4. We observe that utilising personality information (OPN for the speaker in this case) aligns the response closer to the ground truth when compared with the standard BART model.

6.2.6 Human Evaluation

For generative tasks such as response generation, simple reliance on quantitative results proves insufficient, primarily due to the tendency of such metrics, like ROUGE and BLEU scores, to prioritize syntactic similarity over semantic equivalence. Therefore, we perform human evaluation. We conduct a comparative analysis of predictions derived from BART and T5 with and without the incorporation of personality information using PA3. We engage 25 human evaluators[7] who are tasked with assessing a randomly selected set of 50 responses generated by these methods. They assign each response a rating within the range of 1 to 5, considering common human evaluation metrics, including fluency, relevance, coherence, and personality orientation. Detailed definitions for each of these attributes can be found in Appendix A.5.

To monitor the validity of the human evaluations, we calculate Cohen's Kappa (McHugh, 2012) to quantify the inter-annotator agreement between the annotators. The average Kappa score for fluency, coherence, relevancy and personality oriented came out to be 0.83, 0.79, 0.68, and 0.71, respectively. The consolidated results of our human evaluation, shown in Table 5, reflect the averaged ratings across all obtained responses. Evidently, BART, when equipped with personality information using PA3, emerges as the top performer across all metrics.

7 Conclusion

We explored the task of utilising speaker personalities to aid response generation in the domain of code-mixed dialogues. Speaker personalities, from the big five personality traits, are learned in an unsupervised manner an incorporated with dialogue context using a novel fusion mechanism. We leverage a two-level attention mechanism employing context aware and Axial attention approaches to efficiently fuse the personality information with dialogue context. Our experiments demonstrated a notable improvement in response quality and coherence when personality information is fused into the systems. Furthermore, we provided insights into the inferred personality traits and their qualitative connection to response generation.

[7]The evaluators are linguists fluent in English and Hindi with a good grasp of personalized dialogues, aged between 25-30.

8 Limitations

The study does encounter certain limitations that warrant consideration. First, the scarcity of datasets containing multiple dialogues with similar speakers in the code-mixed community limited the study to using a single dataset. While the results show promising outcomes, an investigation with multiple code-mixed datasets can also be beneficial to the community. Additionally, the dataset's source, being from a TV series lacks a real-life-like character development, introducing the possibility of inherent bias. These potential limitation highlights the need for diverse and well-rounded datasets that encompass a variety of conversational scenarios and speaker profiles to ensure the model's applicability across a broader spectrum of code-mixing instances.

9 Ethical Considerations

The study's ethical considerations are well-addressed in several aspects. First, the dataset used in the study is open-sourced and ethically sourced, ensuring that the data collection process adheres to ethical guidelines and data protection regulations. Second, all human annotators and evaluators involved in the research were fairly compensated for their efforts, which is a crucial ethical practice in research involving human participants. Lastly, the study poses no potential concerns related to privacy and consent, as it does not involve the collection or utilization of personal information without explicit permission. These ethical practices help maintain the integrity of the research and ensure that it aligns with ethical standards and principles.

Acknowledgements

The authors acknowledge the support of the ihub-Anubhuti-iiitd Foundation, set up under the NM-ICPS scheme of the DST.

References

Vibhav Agarwal, Pooja Rao, and Dinesh Babu Jayagopi. 2021. Towards code-mixed Hinglish dialogue generation. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 271–280, Online. Association for Computational Linguistics.

Emily Ahn, Cecilia Jimenez, Yulia Tsvetkov, and Alan W Black. 2020. What code-switching strategies are effective in dialog systems? In *Proceedings of the Society for Computation in Linguistics 2020*, pages 254–264.

Firoj Alam and Giuseppe Riccardi. 2014. Fusion of acoustic, linguistic and psycholinguistic features for speaker personality traits recognition. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 955–959. IEEE.

Iqra Ameer, Grigori Sidorov, Helena Gomez-Adorno, and Rao Muhammad Adeel Nawab. 2022. Multi-label emotion classification on code-mixed text: Data and methods. *IEEE Access*, 10:8779–8789.

Suman Banerjee, Nikita Moghe, Siddhartha Arora, and Mitesh M. Khapra. 2018. A dataset for building code-mixed goal oriented conversation systems. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3766–3780, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Anshul Bawa, Pranav Khadpe, Pratik Joshi, Kalika Bali, and Monojit Choudhury. 2020. Do multilingual users prefer chat-bots that code-mix? let's nudge and find out! *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1).

Manjot Bedi, Shivani Kumar, Md Shad Akhtar, and Tanmoy Chakraborty. 2021. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Transactions on Affective Computing*.

Arlin Benjamin Jr. 2020. *Type A/B Personalities*.

Myers Isabel Briggs and Peter B. Myers. 1995. *Gifts Differing : Understanding Personality Type*. Davies-Black Publishing.

Syed Husnain Haider Bukhari, Anusha Zubair, and Muhammad Umair Arshad. 2023. Humor detection in english-urdu code-mixed language. In *2023 3rd International Conference on Artificial Intelligence (ICAI)*, pages 26–31. IEEE.

James N. Butcher and Carolyn L. Williams. 2009. Personality assessment with the mmpi-2: Historical roots, international adaptations, and current challenges. *Applied Psychology: Health and Well-Being*, 1(1):105–135.

Fabio Celli and Bruno Lepri. 2018. Is big five better than mbti? a personality computing challenge using twitter data. In *CLiC-it*.

Guanyi Chen, Yinhe Zheng, and Yupei Du. 2020a. Listener's social identity matters in personalised response generation. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 205–215, Dublin, Ireland. Association for Computational Linguistics.

Guanyi Chen, Yinhe Zheng, and Yupei Du. 2020b. Listener's social identity matters in personalised response generation. *arXiv preprint arXiv:2010.14342*.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explor. Newsl.*, 19(2):25–35.

Paul T Costa and Robert R McCrae. 1992. Normal personality assessment in clinical practice: The neo personality inventory. *Psychological assessment*, 4(1):5.

Paul T Costa Jr and Robert R McCrae. 2008. *The Revised Neo Personality Inventory (neo-pi-r)*. Sage Publications, Inc.

J M Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, 41(1):417–440.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2022. A survey of natural language generation. *ACM Comput. Surv.*, 55(8).

Suman Dowlagar and Radhika Mamidi. 2023. A code-mixed task-oriented dialog dataset for medical domain. *Computer Speech and Language*, 78:101449.

Yifan Fan, Xudong Luo, and Pingping Lin. 2020. A survey of response generation of dialogue systems. *International Journal of Computer and Information Engineering*, 14(12):461–472.

Mauajama Firdaus, Asif Ekbal, and Erik Cambria. 2023. Multitask learning for multilingual intent detection and slot filling in dialogue systems. *Information Fusion*, 91:299–315.

Jennifer Golbeck, Cristina Robles, and Karen Turner. 2011. Predicting personality with social media. In *CHI'11 extended abstracts on human factors in computing systems*, pages 253–262.

Anna Gottardi, Osman Ipek, Giuseppe Castellucci, Shui Hu, Lavina Vaz, Yao Lu, Anju Khatri, Anjali Chadha, Desheng Zhang, Sattvik Sahai, Prerna Dwivedi, Hangjie Shi, Lucy Hu, Andy Huang, Luke Dai, Bofei Yang, Varun Somani, Pankaj Rajan, Ron Rezac, and Yoelle Maarek. 2022. Alexa, let's work together: Introducing the first alexa prize taskbot challenge on conversational task assistance.

Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. 2020. Axial attention in multidimensional transformers.

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

Gabriele Kasper and Johannes Wagner. 2014. Conversation analysis in applied linguistics. *Annual Review of Applied Linguistics*, 34:171–212.

Ankush Khandelwal, Sahil Swami, Syed S. Akhtar, and Manish Shrivastava. 2018. Humor detection in English-Hindi code-mixed social media content : Corpus and baseline system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15):5802–5805.

Shivani Kumar, Sumit Bhatia, Milan Aggarwal, and Tanmoy Chakraborty. 2023a. Dialogue agents 101: A beginner's guide to critical ingredients for designing effective conversational systems.

Shivani Kumar, Atharva Kulkarni, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5956–5968, Dublin, Ireland. Association for Computational Linguistics.

Shivani Kumar, Ishani Mondal, Md Shad Akhtar, and Tanmoy Chakraborty. 2023b. Explaining (sarcastic) utterances to enhance affect understanding in multimodal dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12986–12994.

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, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020a. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020b.

Ro{bert}a: A robustly optimized {bert} pretraining approach.

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020c. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8433–8440.

JM Lucas, F Fernández, J Salazar, J Ferreiros, and R San Segundo. 2009. Managing speaker identity and user profiles in a spoken dialogue system. *Procesamiento del lenguaje natural*, (43):77–84.

Hiren Madhu, Shrey Satapara, Sandip Modha, Thomas Mandl, and Prasenjit Majumder. 2023. Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments. *Expert Systems with Applications*, 215:119342.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Sandip Modha, Thomas Mandl, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Tharindu Ranasinghe, and Marcos Zampieri. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages and conversational hate speech. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 1–3.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

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

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Gopendra Vikram Singh, Mauajama Firdaus, Shambhavi, Shruti Mishra, and Asif Ekbal. 2022. Knowing what to say: Towards knowledge grounded code-mixed response generation for open-domain conversations. *Knowledge-Based Systems*, 249:108900.

Timo Spring, Jacky Casas, Karl Daher, Elena Mugellini, and Omar Abou Khaled. 2019. Empathic response generation in chatbots. In *Proceedings of 4th Swiss Text Analytics Conference (SwissText 2019), 18-19 June 2019, Wintherthur, Switzerland*. 18-19 June 2019.

Naf'an Tarihoran and Iin Ratna Sumirat. 2022. The impact of social media on the use of code mixing by generation z. *International Journal of Interactive Mobile Technologies (iJIM)*, 16(7):54–69.

Mary WJ Tay. 1989. Code switching and code mixing as a communicative strategy in multilingual discourse. *World Englishes*, 8(3):407–417.

William Turnbull. 2003. *Language in action: Psychological models of conversation*. Routledge.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jason Weston, Emily Dinan, and Alexander H Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. *arXiv preprint arXiv:1808.04776*.

Baosong Yang, Jian Li, Derek F. Wong, Lidia S. Chao, Xing Wang, and Zhaopeng Tu. 2019. Context-aware self-attention networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):387–394.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. 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.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018b. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

A.1 Big Five Personality Traits

The widely accepted Big Five Personality Trait Model (Digman, 1990) is a valuable framework for understanding human personality. It consists of five core traits, abbreviated as OCEAN - openness, conscientiousness, extraversion, agreeableness, and neuroticism. These traits provide unique perspectives for character assessment, forming a comprehensive quantitative framework. For detailed definitions and examples of each trait, refer to Table 6.

A.2 Characteristics Descriptions for Dataset Speakers

Drawing insights from Figure 2b, we select the top five frequent speakers, namely Maya, Indravardhan, Sahil, Monisha, and Rosesh, from the extensive MaSaC dataset. These individuals are pivotal for our in-depth analysis. To validate our predicted personalities, human annotators with expertise assess the actual personalities of these speakers. To aid this evaluation, we utilize detailed character descriptions from the Wikipedia page[8] of the show 'Sarabhai v/s Sarabhai'[9], presented in Table 7. Annotators refer to these descriptions when assigning personality traits from the big-five personality traits to each speaker.

A.3 Human Annotations for Evaluating Personality Identification

To validate the RoBERTa model's personality predictions for our top five speakers, we enlisted the input of five human annotators. These annotators, proficient in English and Hindi, were tasked with assigning one of the Big Five personality traits to each speaker based on character descriptions (see Table 7). Their ages range between 25-30. We assessed inter-annotator agreement using the Cohen Kappa method (McHugh, 2012), which yielded an agreement score of 0.78, confirming the reliability of our ground truth.

A.4 Visualisation of Results

In this section, we visualise the ROUGE-1 scores that we obtain for the task of response generation from the standard models without fusing personalities and after fusing personalities using PA3Ḟigure

[8] https://en.wikipedia.org/wiki/Sarabhai_vs_Sarabhai

[9] https://www.imdb.com/title/tt1518542/

Trait	Definition	Example
Openness	This trait reflects a person's willingness to explore new ideas, engage in creative activities, and embrace novel experiences.	Someone high in openness might enjoy trying exotic cuisines, artistic endeavors, and philosophical discussions.
Conscientiousness	Conscientious individuals are organized, goal-oriented, and reliable. They tend to plan ahead and complete tasks with precision.	A conscientious person may meticulously prepare a project schedule and consistently meet deadlines.
Extraversion	Extraversion refers to the degree of sociability, assertiveness, and enthusiasm in an individual.	An extrovert is more likely to enjoy social gatherings, initiate conversations, and thrive in group settings.
Agreeableness	Agreeable individuals are characterized by their empathy, cooperativeness, and willingness to accommodate others.	An agreeable person is more likely to compromise during conflicts and be a supportive friend.
Neuroticism	Neuroticism reflects emotional stability and the tendency to experience negative emotions like anxiety and insecurity.	A highly neurotic person might often worry about various aspects of their life and react strongly to stressors.

Table 6: Definitions and Examples of the Big Five Personality Traits

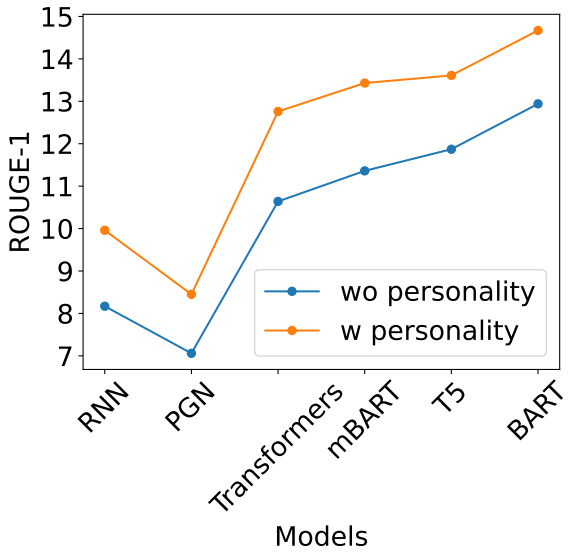


Figure 7: ROUGE-1 score visualisation shows a consistent increase in model performance when personality is infused with dialogue context.

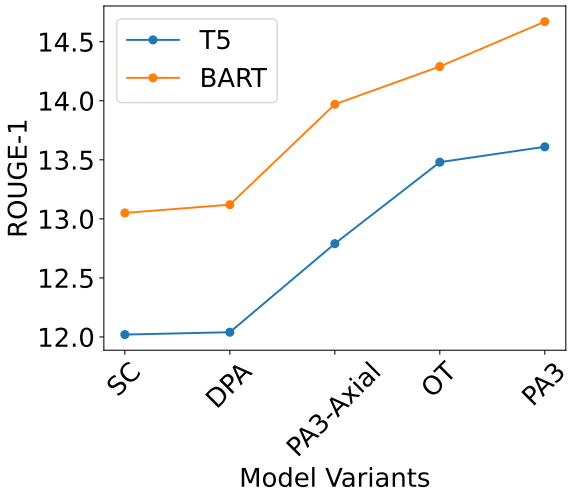


Figure 8: ROUGE-1 score visualisation shows a consistent increase in model performance when we change the fusion method. (Abbr: SC: Simple Concat, DPA: Dot Product Attention, OT: Only Traits, PA3: Personality-Aware Axial Attention).

7 illustrates these findings. It can clearly be observed that there is a consistent increase in the response generation performance when personality is fused into the system for all models. Additionally, we also visualise the increase in performance when we increase the fusion efficiency by ranging the fusion mechanism from simple concat to the proposed PA3 in Figure 8.

A.5 Human Evaluation

For generative tasks like response generation, quantitative metrics alone may not offer a complete evaluation, as they tend to favor syntactic similarity over semantic equivalence. To address this, we utilize human evaluation to provide a more comprehensive assessment. Our approach considers key characteristics to gain a deeper understanding of response quality:

- **Fluency:** This dimension assesses the natural-

ness and readability of the generated text. It focuses on grammar, syntax, and language flow, with higher scores indicating smoother and more linguistically proficient text.

- **Relevance:** The relevance aspect measures how effectively the generated text aligns with the given context or prompt. It evaluates the appropriateness of content in relation to the context, with higher scores signifying a stronger alignment between the response and the context.
- **Coherence:** Coherence evaluation pertains to the logical flow and semantic connection of ideas within the generated text. It ensures that the information is well-structured, logically connected, and readily comprehensible. Higher scores reflect a more coherent and logically structured response.
- **Relevance to Personality:** This specific dimension evaluates whether the generated response is pertinent to the target speaker's personality. It is a crucial element in our evaluation, as it directly relates to the effectiveness of incorporating personality traits into the generated text.

This comprehensive approach offers a nuanced assessment of response generation quality, enhancing our understanding of the system's performance in language, context, and personality capture. See Table 5 for the summarized evaluation results.

A.6 Training System and Hyperparameter Tuning

We mention below the computational framework we use to train our models.

- Description of computing infrastructure used
  - Linux 64 Bit
  - GPU: Tesla-V100 (32510 MiB)
- Trainable parameter: 326368976
- Average runtime: 180 seconds per epoch
- All the results are an average of 3 runs.

After meticulous manual adjustment of hyperparameters, we have identified the ideal parameter configuration. In our exploration of batch sizes, ranging from 2 to 8, we settled on a batch size of 4 due to computational limitations. We chose a learning rate of $5e-6$ with a weight decay of $1e-4$ as lower learning rates led to excessively slow training, while higher rates resulted in erratic learning behavior.

Speaker	Character Description on Wikipedia
Maya	Maya Sarabhai is the female head of the Sarabhai family and runs the family like a pro. Being a snooty upper-class socialite, her daughter-in-law Monisha's middle-class money-saving techniques and unkempt behavior are constant pet peeves for Maya. Her catchphrase is "It's catastrophically middle class!", and she continually uses sarcasm to taunt Monisha and make her see the error of her ways. Whenever she taunts Monisha, depending on the intensity of the taunts, one to three bullet shots are heard in the background, increasing the humor in these situations and portraying her as a verbal bullet. She is constantly after Indravadhan to fix his dietary and cleanliness habits, not much unlike Monisha, and pampers her younger son Rosesh, also making sure he doesn't take a middle-class wife like Sahil. Her son-in-law Dushyant also irritates her by dropping in every time an appliance is damaged.
Indravardhan	Indravadhan Sarabhai a.k.a. Indu, is an ex-director of a multinational company. He retired early to take care of the children and help Maya work as a social worker. He is always in conflict with his youngest son, Rosesh, he also jokes with Maya, pretending to hate her but actually loving her dearly as portrayed in various episodes. He constantly picks on Maya and Rosesh, always siding with Monisha in case of a tiff between her and Maya, and constantly tries to create conflicts between them. He notoriously ignites most of the quarrels in the family and then takes the seat in the audience, enjoying himself. He is irritated by his brother-in-law Madhusudan Bhai and his "hain?", as well as Dushyant, his son-in-law. He is the jester in the family.
Sahil	Sahil Sarabhai is a cosmetologist. He is the eldest child, and arguably the most normal one in his otherwise eccentric family. He is soft, calm, wise and noble, and is constantly trying to resolve conflicts in his family, between Maya and Monisha, Maya and Indravadhan and Rosesh. He often gets sandwiched between his mother and his wife and tries not to hurt anyone. He avoids conflicts but loves making fun of his younger brother Rosesh, similar to Indravadhan.
Monisha	Monisha Sarabhai is a middle class, Punjabi girl from Noida and now the daughter-in-law of the Sarabhai's. She rarely cleans the house and is always lazing around watching daily soaps on television. She develops a dramatic nature from these shows and always ends up saying threatening Sahil with leaving the house after every argument with Maya. Her passion is to save money, come what may. She is always at loggerheads with Maya for her thrifty ways. Her father-in-law always supports her, while Sahil is torn between the two. Despite being careless, Monisha is an honest, innocent, and loving woman. Manisha was named 'Monisha' by Maya as she found the name Manisha 'too middle-class'.
Rosesh	Rosesh Sarabhai is the youngest child of Maya and Indravadhan. He is a theatre artist, an aspiring actor, and a so-called poet. He is Maya's favorite and she pampers him a lot. He wants to become an actor and his mother Maya supports him the most. Maya is the only member of the Sarabhai family who approves of and appreciates his absurd poetry and acting skills. He has a love-hate relationship with Indravadhan as he is always the target of his jokes and pranks. He always seconds his momma even if he doesn't feel like it. He has a peculiar and amusing voice, and his poems are always bad but funny.

Table 7: Character definition as present on Wikipedia of the most frequent five speakers in MaSaC dataset.