RECAP: Retrieval-Enhanced Context-Aware Prefix Encoder for Personalized Dialogue Response Generation

Shuai Liu **Hyundong J. Cho** **Marjorie Freedman**
Xuezhe Ma **Jonathan May**
Information Sciences Institute
University of Southern California
{liushuai, jcho, mrf, xuezhema, jonmay}@isi.edu

Abstract

Endowing chatbots with a consistent persona is essential to an engaging conversation, yet it remains an unresolved challenge. In this work, we propose a new retrieval-enhanced approach for personalized response generation. Specifically, we design a hierarchical transformer retriever trained on dialogue domain data to perform personalized retrieval and a context-aware prefix encoder that fuses the retrieved information to the decoder more effectively. Extensive experiments on a real-world dataset demonstrate the effectiveness of our model at generating more fluent and personalized responses. We quantitatively evaluate our model's performance under a suite of human and automatic metrics and find it to be superior compared to state-of-the-art baselines on English Reddit conversations.[1]

1 Introduction

As tremendous successes have been achieved on open-domain dialogue generation (Zhang et al., 2020b; Cho and May, 2020; Roller et al., 2021; Shuster et al., 2022), personalized dialogue models have started to draw attention because of their ability to generate consistent and engaging conversations and their potential time-saving utility in on-message predictive generation (Wu et al., 2021; Ma et al., 2021b; Zhong et al., 2022). To generate persona-consistent responses, these models condition on not only dialogue context but user personas, which can be either explicitly given or implicitly learned from the user conversations. Early works mostly focus on modeling explicit personas (Zhang et al., 2018; Zheng et al., 2019; Song et al., 2019, 2021). These methods rely on dialogue data paired with user traits, profiles or persona description sentences, which are difficult to collect in practice. Moreover, explicit personas usually only contain a few user traits (e.g. age, gender, and location) or a few profile sentences, so the amount of information carried with them is limited, which restricts the models' capability to capture and then express more nuanced personalization. Later works develop methods for automatically extracting personas (Mazaré et al., 2018; Wu et al., 2020), in order to help improve content diversity, as compared to that seen in explicit personas. However, these extraction methods still cannot fully use all information from user history conversations.

Recent works address these issues by incorporating user dialogue history as their implicit profiles (Wu et al., 2021; Ma et al., 2021b; Zhong et al., 2022). These methods generate personalized responses in two phases, retrieving relevant conversations from the user history and fusing the retrieved information to the generator. In the first phase, these methods retrieve a subset of conversations from a user's conversation history (Ma et al., 2021b; Zhong et al., 2022). In the second phase, the retrieved conversations are fused into a decoder by manipulating output logits (Ma et al., 2021b; Wu et al., 2021) or by adding prompt tokens to the input (Zhong et al., 2022). Even though the implicit profile approach is shown to be the most robust and scalable among all approaches on real-world datasets, it still has some potential weaknesses. Approaches to the retrieval phase include using recent conversations (Ma et al., 2021b) or using conversations based on current context similarity, according to an out-of-domain model (Wu et al., 2021; Zhong et al., 2022). These approaches have the potential to lose important personal information, which may lead to unexpected behavior and poorly motivated retrieval. In the fusion phase, neither output logit manipulation nor input token prompting may fully leverage the capability of the pre-trained decoder. In this work, we focus on the implicit user profile approach, but specifically address the weaknesses in both retrieval and fusion phases.

[1] Our code and data are publicly available at https://github.com/isi-nlp/RECAP.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8404–8419
July 9-14, 2023 ©2023 Association for Computational Linguistics

We present RECAP, a **R**etrieval-**E**nhanced **C**ontext-**A**ware **P**refix encoder for personalized dialogue response generation. Similar to other implicit user profile methods, our model is based on a retrieval-fusion approach, which first retrieves persona-relevant information, and then fuses it with conversation context at decode-time. Unlike previous work, which does not take the purpose of the retrieval task into consideration, our hierarchical transtormer retriever is trained specifically to retrieve information that will best communicate a user's persona. Unlike previous work, which approaches fusion by concatenating retrieved information at the input level (Zhong et al., 2022) or manipulating logits at the output level (Ma et al., 2021b; Wu et al., 2021), we adopt a continuous pre-layer prefix approach (Li and Liang, 2021; Liu et al., 2022), along with a two-step cross-attention projection (Humeau et al., 2020; Ma et al., 2021a), both of which have been shown to be beneficial. These novelties result in a better personalized dialogue model.

Our main contributions in this work are:

- We design a hierarchical transformer retriever that can perform personalized history retrieval based on different target users using their history conversations.

- We design a context-aware prefix encoder that can encode context-relevant information from user histories and fuse the information to the generator effectively through the prefix.

- The two modules combined achieve state-of-the-art performance on personalized dialogue response generation for English Reddit conversations by generating fluent and personalized responses for unseen users, as shown in automatic and human evaluations.

2 Methodology

In this section, we formalize the personalized dialogue response generation task and introduce our proposed RECAP method.

2.1 Task Definition

Our goal is to build a personalized dialogue model that generates persona-consistent responses with a target user's *history* of conversations. Formally, we have a set of users $\mathcal{U}$ but will henceforth assume user $u \in \mathcal{U}$ is the *target user*, that is, the user we wish to personalize, User u's history is represented as a set of context-response pairs $\mathcal{H}_u = \{(\mathbf{c}_1^u, \mathbf{r}_1^u), \cdots, (\mathbf{c}_T^u, \mathbf{r}_T^u)\}$, where a *context* $\mathbf{c}_t^u$ is a sequence of one or more turns that starts at the beginning of a conversation and ends with the turn immediately before the single turn *response* $\mathbf{r}_t^u$, which is by definition authored by u.[2] Given some *current* (context, response) pair $(\mathbf{c}^u, \mathbf{r}^u) \notin \mathcal{H}_u$, we seek to maximize

$$p(\mathbf{r}^u|\mathbf{c}^u, \mathcal{H}_u) = \prod_{i=1}^{|\mathbf{r}^u|} p(r_i^u|\mathbf{c}^u, \mathbf{r}_{<i}^u, \mathcal{H}_u) \qquad (1)$$

where $\mathbf{r}_{<i}^u$ represents tokens preceding token r_i^u in $\mathbf{r}^u$.[3]

2.2 Model Overview

RECAP consists of two main modules: a retrieval module (RE), which selects user history responses, and a context-aware prefix encoder (CAP), which converts the selected responses into a suitable dense prefix. The prefix, when prepended to our transformer decoder's intermediate states as in Liu et al. (2022), yields personalized generation. In the following we describe each of RE and CAP in more detail.

2.3 Retrieval Module (RE)

We follow a standard bi-encoder retrieval approach as in Wu et al. (2018): a dense representation is formed for each of u's candidate dialogue turns, which we regard as *documents*, as well as for a *query* representing the context of the conversation for which a turn will be generated. The set of documents that is closest (as measured by cosine) to the query is returned.

Using such retrieval methods for dialogue personalization is not novel (Isbell et al., 2006; Zhang et al., 2018; Wu et al., 2021; Ma et al., 2021b; Zhong et al., 2022), but previous work simply either retrieves the user's most recent turns (Ma et al., 2021b) or simply queries for turns based on the similarity to the current conversation or a predicted topic (Zhong et al., 2022). Our retrieval model, by contrast, forms a query based on an *a priori* predicted next turn given the current context and user history.

[2] Some, but not all, of the turns in $\mathbf{c}_t^u$ may have been authored by u.

[3] Unless otherwise noted, we drop the superscript henceforth.

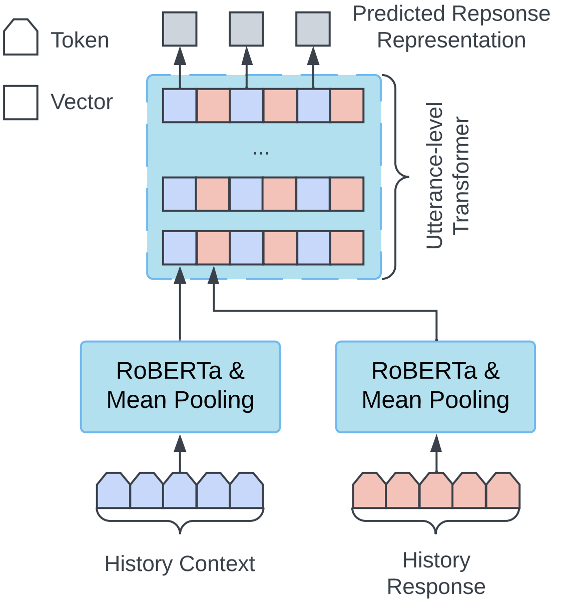

Figure 1: The architecture overview of the retrieval module (RE) based on hierarchical transformer.

The representation of the predicted next turn is learned based on not only the context of the immediate conversation, but also based on every other conversation known to involve u, forming an ersatz persona. This can present a problem. Prolific users can potentially have had a long history of conversations from which to choose to represent an implicit persona, and it is unrealistic to use this entire history when generating a single response. Furthermore, existing works have shown that using a small subset of the history is more beneficial to implicit persona-based generation, as doing so reduces noise and computation time (Wu et al., 2021; Zhong et al., 2022). Inspired by the hierarchical dialogue model (Serban et al., 2015, 2016) and the hierarchical transformer (Pappagari et al., 2019; Zhang et al., 2019), we build a response prediction model that takes as input a user's available history and the current dialogue context in an efficient manner. The hierarchical architecture is shown in Figure 1.

Specifically, we first concatenate all turns in the oldest history context, $\mathbf{c}_1$, encode them with a pretrained RoBERTa model (Liu et al., 2019), and form a fixed-length representation from the mean of the last hidden states corresponding to each token. To this we add positional embedding $\mathbf{p}_1$ indicating it is the first history context known, and utterance type embedding $\mathbf{y}_c$ indicating a context representation. We then do the same for the analogous $\mathbf{r}_1$, and in turn for all other context and response pairs in the history. In general, the inputs to the next level of the hierarchical transformer are formed as follows:

$$\mathbf{e}_{\mathbf{c}_t} = \text{mean}(\text{RoBERTa}(\mathbf{c}_t)) + \mathbf{p}_t + \mathbf{y}_c \quad (2)$$

$$\mathbf{e}_{\mathbf{r}_t} = \text{mean}(\text{RoBERTa}(\mathbf{r}_t)) + \mathbf{p}_t + \mathbf{y}_r \quad (3)$$

We term these *utterance-level* embeddings and pass $[\mathbf{e}_{\mathbf{c}_1}, \mathbf{e}_{\mathbf{r}_1}, \cdots, \mathbf{e}_{\mathbf{c}_T}, \mathbf{e}_{\mathbf{r}_T}]$ to an utterance-level transformer, yielding the sequence of hidden representations $[\mathbf{h}_{\mathbf{c}_1}, \mathbf{h}_{\mathbf{r}_1}, \cdots, \mathbf{h}_{\mathbf{c}_T}, \mathbf{h}_{\mathbf{r}_T}]$. We train the transformer to predict the ground truth response representation and minimize cosine similarity between the predicted and ground truth representation, i.e.

$$\mathcal{L} = \sum_{t=1}^{T} 1 - \frac{\mathbf{h}_{\mathbf{r}_t} \cdot \mathbf{g}_{\mathbf{r}_t}}{|\mathbf{h}_{\mathbf{r}_t}||\mathbf{g}_{\mathbf{r}_t}|} \quad (4)$$

where $\mathbf{g}_{\mathbf{r}_t}$ is the ground-truth representation of the response at time t encoded by an off-the-shelf sentence transformer. A causal mask is applied during training to prevent attention to future utterances.

This general architecture can be specialized by changing the underlying pretrained RoBERTa model used for token-level embedding. In particular, we consider two types of representation: (1) a *style* representation, which we obtain by encoding histories with an off-the-shelf content-independent style representation model (Wegmann et al., 2022),[4] and (2) a representation encoded by a sentence transformer (Reimers and Gurevych, 2019)[5] which we term a *semantic* representation, to contrast the style representation.

For retrieval, we first predict the style and semantic representations of the response to be generated and retrieve the appropriately embedded history responses whose style/semantic representations are the most similar to the predicted style/semantic representations. The retrieved history responses are then passed to the context-aware prefix encoder (Section 2.4) for further encoding.

2.4 Context-Aware Prefix Encoder (CAP)

The purpose of the CAP module is to project the history responses retrieved by RE (Section 2.3) onto a fixed-length prefix vector. This vector is then prepended to the transformer decoder hidden states as a prefix. The architecture is illustrated

[4] https://huggingface.co/AnnaWegmann/Style-Embedding

[5] https://huggingface.co/sentence-transformers/all-distilroberta-v1

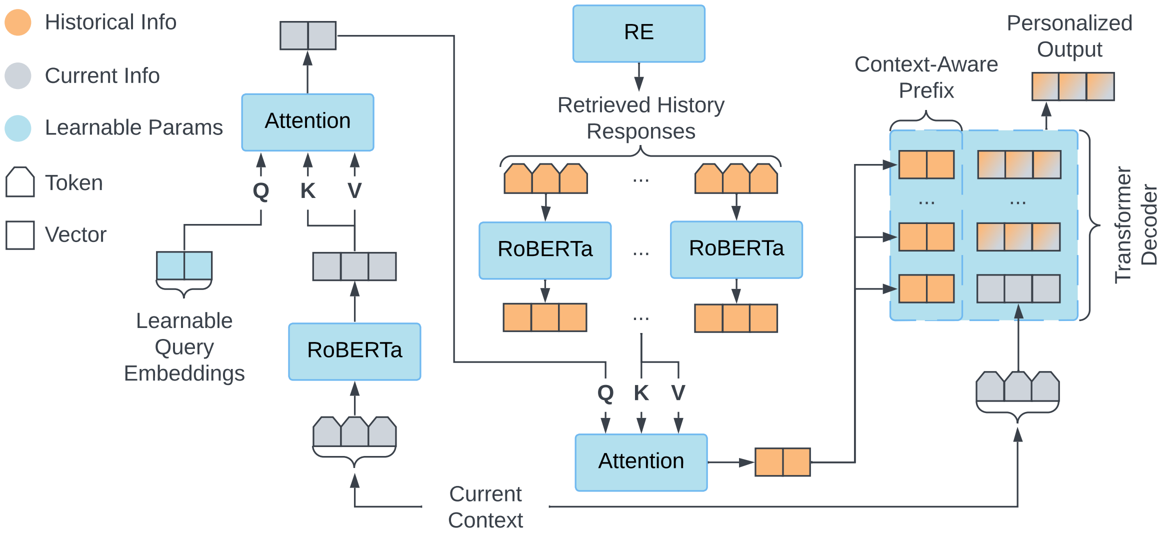

Figure 2: The architecture overview of the context-aware prefix encoder (CAP) and the decoder generator.

in Figure 2. CAP first encodes the current dialogue context and each of the retrieved responses to continuous representations with a pre-trained RoBERTa encoder (Liu et al., 2019).[6] We use the same method as Liu et al. (2021) to add positional embeddings to user history response representations:

$$\mathbf{C} = \mathrm{RoBERTa}(\mathbf{c}) \tag{5}$$

$$\mathbf{H}_i = \mathrm{RoBERTa}(\mathbf{h}_i) + \mathbf{q}_i \tag{6}$$

where $\mathbf{c}$ is the current context, $\mathbf{h}_i$ is the i-th retrieved history response, $\mathbf{C}$ and $\mathbf{H}_i$ are the last hidden states from the application of RoBERTa to the every token position of $\mathbf{c}$ and the i-th retrieved history response, respectively, and $\mathbf{q}_i$ is a history positional embedding for retrieved history i. All $\mathbf{H}_i$'s are then concatenated to a long vector sequence $\mathbf{H} = [\mathbf{H}_1; \cdots; \mathbf{H}_{t-1}]$.

Inspired by the cross-attention context projection operation (Humeau et al., 2020; Ma et al., 2021a), CAP projects the long vector sequence $\mathbf{H}$ onto a short fixed-length prefix with two cross-attention operations, which we denote as $\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ to indicate which information source is used as query, key, and value, respectively.[7] Separate query, key, and value matrices are learned for each of the two operations.

The first cross-attention operation queries $\mathbf{C}$ with learnable query embeddings $\mathbf{E} \in \mathbb{R}^{Nd}$ and projects to fixed-length representation $\mathbf{P}_c \in \mathbb{R}^{Nd}$; N is a chosen hyperparameter and d is RoBERTa's token embedding dimension:

$$\mathbf{P}_c = \mathrm{Attn}(\mathbf{E}, \mathbf{C}, \mathbf{C}) \tag{7}$$

Then, the second operation queries the user history representations $\mathbf{H}$ with $\mathbf{P}_c$ to obtain the fixed-length context-aware user history representations $\mathbf{P}_h \in \mathbb{R}^{Nd}$:

$$\mathbf{P}_h = \mathrm{Attn}(\mathbf{P}_c, \mathbf{H}, \mathbf{H}) \tag{8}$$

Finally, similar to a memory vector projection (Li et al., 2020), $\mathbf{P}_h$ is projected onto $\mathbb{R}^{LNd}$ with a linear layer and then separated into L d-dimensional vector sequences with length N, corresponding to the L layers in the transformer decoder. Each of these sequences is then prepended to the transformer decoder hidden state in the analogous layer.

2.5 Generator

We use the pre-trained DialoGPT (Zhang et al., 2020b)[8] as the generator. Personalized information is fused to the generation process through the prefix vectors encoded by CAP, as described in Section 2.4. We further train the parameters in DialoGPT together with the CAP module to maximize the (log of the) objective in Equation 1.

3 Experiments

3.1 Dataset

We extract a personalized conversation dataset from Reddit on pushshift.io (Baumgartner et al., 2020).[9]

[6] https://huggingface.co/docs/transformers/model_doc/roberta

[7] The information source for key and value is always the same, as is typical.

[8] https://huggingface.co/docs/transformers/model_doc/dialogpt

[9] https://files.pushshift.io/reddit/

We choose only from the conversations from August 2019 to June 2021 to avoid test data leakage when using pre-trained models. Each sample in the dataset consists of three entries: user name, context (i.e preceding turns in the conversation), and response. Since the total number of samples is very large, we randomly select 115,000 users. For each selected user, we keep only the 10 most recent samples to train the generator and the 100 most recent samples as history conversations to retrieve from. Unlike existing works (Wu et al., 2021; Ma et al., 2021b; Zhong et al., 2022) that use the same users for training, validation, and test, we partition the dataset by user. In this way, we can test the model's ability to generalize to unknown users. Specifically, we select 100,000, 5,000, and 10,000 distinct users for each of training, validation, and test, respectively.

3.2 Baseline Models

We compare our model with four baseline models including the state-of-the-art personalized dialogue model.

- **DialoGPT:** A large-scale pre-trained dialogue response generation model trained on Reddit conversations (Zhang et al., 2020b).

- **DialoGPT w/ history responses:** We directly prepend retrieved history responses to the DialoGPT input (i.e. dialogue context).

- **DHAP:**[10] A model that generates personalized responses by building a dynamic context-aware user profile representation from user history conversations and then employing a personalized decoder with a copy mechanism (Ma et al., 2021b). We enhance DHAP with pre-trained transformers for fair comparison to our model.

- **MSP:**[11] The state-of-the-art personalized dialogue model. MSP generates personalized responses by prepending selected tokens directly to the DialoGPT input. The tokens are selected by a three-step hierarchical refiner (Zhong et al., 2022).

3.3 Implementation Details

Our implementation is based on HuggingFace's Transformers (Wolf et al., 2020)[12] and Sentence Transformer (Reimers and Gurevych, 2019)[13] codebases. We experiment with different settings and hyperparameters; the ones that work the best are discussed below. We initialize all encoders from the pre-trained RoBERTa-base model (Liu et al., 2019)[14] and initialize all decoders from the pre-trained DialoGPT-small model (Zhang et al., 2020b). RoBERTa's embedding dimension, d, is 768, and N, the prefix length, is set to 30 to align with the prompt length used in MSP. The two projection attentions in CAP are both single-head attentions. The number of history responses are 10, and for the models without a retrieval module (DialoGPT + history, DHAP, and CAP), the 10 most recent history responses are used. The utterance-level transformer has 768 hidden dimension, six layers, and a 12-head self-attention in each layer. We train all models using the AdamW optimizer (Loshchilov and Hutter, 2019) with learning rate 5e-5 and linear learning rate schedule for 10 epochs. The best models are chosen based on the validation perplexity. For generation, we use nucleus (top-p) sampling (Holtzman et al., 2020) with $p = 0.8$.

	# Params	Training Time (hr)
DialoGPT	124M	8
DialoGPT + history	124M	37
DHAP	431M	25
MSP	437M	15
RE	198M	13
CAP	269M	22

Table 1: Total number of parameters and training time (on 2 × A40 GPUs) for all neural models.

All models including the baseline models are trained on 2 × A40 GPUs with half precision training. The total number of parameters and training time are shown in Table 1. For MSP, the numbers are sums over all sub-modules (i.e. three refiners and DialoGPT generator). CAP includes both the context-aware prefix encoder and the DialoGPT generator.

3.4 Evaluation Metrics

3.4.1 Automatic Evaluation

In this section, we discuss the automatic evaluation metrics we use to evaluate all models. We group

[10] https://github.com/zhengyima/DHAP

[11] https://github.com/bangbangbang12315/MSP

[12] https://huggingface.co/docs/transformers

[13] https://www.sbert.net

[14] We specifically use RoBERTa rather than some other core representation for compatibility purposes; RoBERTa and DialoGPT share vocabulary, which is required by two of the baseline models, DHAP and MSP.

them into four categories. The first two categories measure the general performance of the models, while the last two measure the personalization ability.

Perplexity and Token-overlap Metrics We first evaluate the performance of our model with several of the most commonly used automatic evaluation metrics for dialogue response generation, including perplexity, BLEU-1 and -2 (Papineni et al., 2002), ROUGE-L (Lin and Och, 2004), and METEOR (Banerjee and Lavie, 2005). Perplexity evaluates how well a language model predicts the sample. Lower perplexity means the model can generate a more fluent response and generalizes better (Blei et al., 2003). The other metrics are word or n-gram overlap metrics with a reference utterance. A higher score means a higher similarity between the generated text and the ground-truth text since they have more words or phrases in common.

Learning-based Metrics Many learning-based metrics backed with pre-trained models have been developed. They are shown to be more robust and correlate better with human judgement than token overlap metrics, though issues have been raised regarding their inherent biases (Gowda et al., 2021). In this work, we select two of the most popular learning-based metrics: BERTScore (Zhang et al., 2020a) and BLEURT (Sellam et al., 2020). These two methods also measure how similar a candidate response is to the reference response; higher scores mean higher similarity.

Style Metrics To measure the models' ability to capture personal writing styles, we employ a pre-trained style representation model (Wegmann et al., 2022) for evaluation. We form two metrics based on the style model: (1) embedding similarity and (2) contrastive authorship verification (CAV) accuracy. Embedding similarity is simply the cosine similarity between the style embedding of the generated response and that of the ground-truth response. For CAV accuracy, we construct a domain-controlled (Wegmann et al., 2022) dataset with response triplets built from a generated response anchor and a pair of positive/negative ground-truth responses. The positive example and negative example are from the same author as the anchor and a randomly sampled author, respectively. With domain control, we only choose the negative example from the same subreddit as the anchor and the positive example from a different subreddit. To evaluate, we calculate the percentage of the triplets for which the style model judges the generated anchor response to be more similar to the positive ground-truth response than to the (randomly sampled) negative example. For both style metrics, a higher score indicates better personalization.

	Train	Validation	Test
Age	1378	344	585
Gender	1823	455	784
MBTI	5542	1385	2131

Table 2: Number of train, validation, and test users in the Pandora dataset.

Personal Traits Metrics A good personalized response model should also be able to reflect the personal traits of the target user. Therefore, a personal traits classifier is also used as a evaluation method in previous works. Zheng et al. (2019) evaluate their model on traits of age, gender, and location, while Xing and Fernández (2018) proposed a evaluation method based on a personality classifier. In this work, we select three personal traits for evaluation: age, gender,[15] and Myers-Briggs Type Indicator (MBTI) (Briggs-Myers and Myers, 1995).[16] We train a model for each trait on the PANDORA dataset (Gjurković et al., 2021),[17] then attempt to determine traits based on generated responses. A good personalized model should generate output that allows a trait classifier to guess traits about as well as it can when given actual responses.

For age, we train a linear regression model and report the Pearson correlation coefficient (Benesty et al., 2009) between the predicted age and the ground-truth age. For gender and the four MBTI categories, we train a logistic regression model for each and report the classification F1 score. For all three metrics, higher scores are better. We se-

[15] Due to data limitation issues, we simplify gender identity as a binary.

[16] We acknowledge there is extensive criticism of MBTI in the psychology community (Capraro and Capraro, 2002; Pittenger, 2005) and do not address the validity of fixed personality types, of MBTI as an approach to determining types, or even of the general predictability of MBTI labels given dialogue text in this work. However, following other work (Kishima et al., 2021; Sang et al., 2022), we include this metric, which indicates the degree to which the generated responses for a user are as predictive of self-declared MBTI labels as the user's actual responses. The MBTI metric is not the only metric we use, and its inclusion is meant to accompany the others to indicate a consistent trend toward user-like generation.

[17] https://psy.takelab.fer.hr/datasets/all/pandora

lect users for personal traits evaluation from the PANDORA dataset, which is independent from our Reddit test set. Table 2 shows the statistics for the PANDORA dataset.

The input features, for all models, are the most frequent 40,000 TF-IDF weighted 1-3 word ngrams. The logistic regression models use the L-BFGS solver. The best C value (i.e. inverse of regularization strength) for gender and the four MBTI categories (i.e. I/E, S/N, T/F, J/P) are 10, 1, 10, 50, and 0.1, respectively. All models can be trained within 20 seconds on a single CPU with scikit-learn (Pedregosa et al., 2011).[18]

3.4.2 Manual Evaluation

We also conduct a manual evaluation. We randomly sample 100 examples from the test set for all models and hire two well-educated volunteer annotators (one of the authors and a friend of one of the authors). The annotators evaluate responses on three criteria: fluency, coherency, and persona consistency. All criteria are scaled from 1 to 3 (from disagree to strongly agree). First, for fluency, we only show the response and ask "is the response overall readable and fluent?" Then, for coherency, we also show the preceding turns in the conversation and ask "does the response serve as a valid continuation of the preceding conversation?" Finally, for persona consistency, we show five ground-truth responses written by the target author and ask "does the response seem like it would have been written by the author of the given texts?"

4 Results

In this section, we discuss the experimental results and further analysis. Due to limited time and computational resources, we only report the results from a single run, and run statistical significance tests.[19]

4.1 Automatic Evaluation Results

Table 3 shows the automatic evaluation results for all models on selected metrics. For conciseness, we only show a representative or aggregated metric for similar metrics, but the full results are shown in Appendix C and are generally consistent with the representative results shown in Table 3. In nearly all cases, the top two results in all automatic metrics are from our models. Without the retriever, the CAP model already outperforms the baseline models on most automatic metrics. With the retrieval enhancement, the RECAP models achieve better scores on most automatic metrics. Specifically, with style retrieval enhancement, the RECAP model obtains better style embedding similarity, CAV accuracy, and average MBTI F1 score, which indicates better performance at reflecting the target author's writing style. With semantic retrieval enhancement, the RECAP model achieves the best scores on token-overlap metrics and learning-based metrics, which indicates it can generate responses that are more similar to the ground-truth. Moreover, combining two enhancement methods by mixing half retrieved history responses from each retriever also combines the strength of the two RECAP models. Even though the combination also weakens the improvements, we can still see that the RECAP-mixed model is at least the second best on all metrics on the Reddit dataset.

4.2 Human Evaluation Results

Table 4 shows the human evaluation results. Cohen's κ and Krippendorff's α between the two annotators are $\kappa = 0.617$ and $\alpha = 0.687$, respectively. The κ shows a substantial agreement between the two annotators, and the α indicates that a tentative conclusion could be drawn from the human evaluation results (Antoine et al., 2014). Even though there are some minor inconsistencies, both human annotation results and automatic evaluation results on the Reddit dataset agree on the top two models, which are RECAP-semantic and RECAP-mixed for general response quality, and RECAP-style and RECAP-mixed for style/persona metrics. Furthermore, RECAP-mixed is the overall second best model under human evaluation.

4.3 Style Consistency Analysis

Even though the automatic and human metrics give us a general idea of the model performance, these scores are not very interpretable. To further understand style consistency beyond the metric scores, we conduct a case analysis similar to Wegmann et al. (2022) by inspecting whether the models can capture some aspects of writing style. Specifically, we select three aspects mentioned by Wegmann et al.: last punctuation (i.e. whether the response ends with a punctuation mark), contraction spelling (i.e. whether the response uses "n't" or "nt" in contractions like "didn't"), and casing (i.e. whether the response is all lowercased). For each aspect, we cal-

[18] `https://scikit-learn.org/stable`

[19] Please refer to Appendix B for details.

Model	Reddit					PANDORA		
		Token-overlap	Learning-based	Style Metric		Demographic		MBTI
	PPL↓	ROUGE-L↑	BLEURT↑	Embed Sim↑	CAV Acc↑	Age↑	Gender↑	Average↑
DialoGPT	31.25‡	8.49‡	0.2421‡	22.15‡	51.14‡	0.0425‡	0.6103‡	0.4992‡
DialoGPT + history	29.66‡	9.84‡	0.2482‡	40.66‡	64.02	0.1008‡	0.6763‡	0.5130‡
DHAP	29.99‡	9.73‡	0.2511‡	37.14‡	61.53‡	0.0829‡	0.6653‡	0.5119‡
MSP	30.47‡	9.36‡	0.2453‡	34.72‡	59.61‡	0.1226‡	0.6816‡	0.5019‡
CAP	**29.44**	10.09‡	0.2534‡	40.38‡	63.71‡	**0.2051**	0.7077†	0.5167†
RECAP-style	29.54‡	10.05‡	0.2525‡	**41.40**	64.14	0.1822	0.7001†	0.5265
RECAP-semantic	29.50‡	**10.33**	**0.2749**	39.65‡	64.00	0.1699†	**0.7303**	**0.5276**
RECAP-mixed	29.47†	10.27†	0.2557†	40.74†	**64.17**	0.1392‡	0.6962†	0.5242
Ground-truth	-	-	-	-	66.20	0.2617	0.7477	0.5257

Table 3: The automatic evaluation results on Reddit and PANDORA datasets with selected metrics.The best and second best results in each column are shown in **bold** and underline, respectively. Scores for ground-truth are not available for metrics calculated based on ground-truth, and they are shown by "-". "†" and "‡" indicates statistically significant difference for $p < 0.05$, between the best or the top two models, respectively, determined by t-test.

Model	Fluency↑	Coherency↑	Persona↑
DialoGPT	2.75‡	2.27‡	1.58‡
DialoGPT + history	2.77	2.28‡	1.84‡
DHAP	2.72‡	2.28†	1.76‡
MSP	2.73‡	2.29‡	1.85‡
CAP	2.72‡	2.31	1.90‡
RECAP-style	2.77	2.28‡	**2.03**
RECAP-semantic	**2.80**	**2.35**	1.92‡
RECAP-mixed	2.79	2.33	2.00
Ground-truth	2.84	2.40	2.47

Table 4: The human evaluation results on the Reddit dataset. The best and second best results in each column are shown in **bold** and underline. "†" and "‡" indicates statistically significant difference for $p < 0.05$, between the best or the top two models, respectively, determined by t-test.

Model	Punc.↑	Cont.↑	Casing↑
DialoGPT	0.3538	0.3822	0.3300
DialoGPT + history	0.4117	0.4333	0.4180
DHAP	0.3829	0.4121	0.3831
MSP	0.3795	0.4064	0.3788
CAP	0.4144	0.4415	0.4183
RECAP-style	0.4112	0.4403	0.4172
RECAP-semantic	**0.4232**	**0.4520**	**0.4248**
RECAP-mixed	0.4178	0.4451	0.4195

Table 5: Style consistency analysis results for all models. The best and second best results in each column are shown in **bold** and underline.

culate the percentage of generated responses that match the ground-truth style. Table 5 shows the results, which indicate that most of our models can capture all three selected aspects more effectively than the baseline models. The lone exception is the RECAP-style model which is slightly worse than the DialoGPT + history model on last punctuation and casing aspects.

5 Related Work

Personalized Dialogue Model Recent works on personalized response generation mainly fall into three categories: (1) those that personalize the response with a user-specific embedding learned during training (Li et al., 2016; Chan et al., 2019), (2) those that personalize the response with explicit user profiles or persona description sentences (Zhang et al., 2018; Zheng et al., 2019; Song et al., 2019, 2021), and (3) those that personalize the response with an implicit user persona extracted from user history conversations (Bak and Oh, 2019; Wu et al., 2021; Ma et al., 2021b; Zhong et al., 2022). User-specific embeddings are shown to be ineffective and hard to generalize to unseen users since the embeddings need to be learned during training (Zhong et al., 2022). Explicit user profiles and personas require manual data collection, which is very hard to scale up in practice and is often not available in deployed scenarios. Recent works (Ma et al., 2021b; Zhong et al., 2022) show strong scalability and robustness of the implicit user persona based method, and for that reason our work also focuses on this method.

The state-of-the-art implicit user persona method MSP (Zhong et al., 2022) incorporates a three-step hierarchical refiner to select informative tokens from relevant history responses from similar users, and the selected tokens are then prepended to the transformer deocder input as a prompt to personalize the generation process. However, their response selection module is trained on a news dataset that, because of domain divergence, may lead to sub-optimal retrieval performance on the intended dialogue task. Further, the hard discrete token selection module employed in MSP may be further improved by instead using a continuous prompt/prefix. Therefore, inspired by the hierarchical dialogue model and hierarchical transformer, we develop a personalized retrieval model that can use all user history conversations. As suggested by Dudy et al. (2021), we develop a personalized generator with a prefix mechanism similar to that used in Li and Liang (2021) and Liu et al. (2022), but instead of learning the prefix during training, we train a prefix encoder to dynamically encode a personalized prefix with user history responses so that the model can be easily generalized to unseen users without further training.

Our model has two main differences from MSP: (1) we use a personalized response retriever trained on dialogue domain instead of a non-personalized retriever trained on distant news domain. (2) we use a dynamically encoded continuous prefix to fuse personalized retrieved responses to the generator rather than a discrete token prompt.

Hierarchical Transformer Hierarchical transformers model long documents with a sentence-level transformer on top of a regular token-level transformer. The token-level transformer represents each sentence as a single vector embedding, and the embedding vectors of all sentences in the document are concatenated together and fed to the sentence-level transformer as input. These models are shown to be effective for long text classification (Pappagari et al., 2019) and summarization (Zhang et al., 2019). Our retrieval module uses a similar hierarchical transformer for response utterance-level embedding prediction, but differs in tasks and training strategy. Our retrieval module is trained on a generative next response prediction task with utterance-level causal masks.

6 Conclusion

In this work, we introduce RECAP, a personalized dialogue model, which generates responses in a retrieval augmentation manner. Unlike retrievers used in previous works, the hierarchical transformer retriever can perform personalized retrieval using user history responses. The context-aware encoder can encode and preserve the most useful information from the retrieved responses and fuse the information to a regular transformer decoder through continuous prefix vectors. Extensive experiments confirm that our model is capable of generating fluent, coherent, and personalized responses.

Ethical Issues

Like most data-driven dialogue models, our model is trained on a large-scale naturally-occuring dataset, the Pushshift Reddit dataset (Baumgartner et al., 2020), which may contain biased and offensive content. To preserve persona and personal writing style as much as possible, we did not filter out conversations with this content. To avoid potentially unethical responses in real-world usage, we suggest filtering out the data with unethical content before training or applying a post-generation filter for the offensive responses.

Even though our model is intended to generate personalized responses for only personal usage (e.g. personal virtual assistant), we realize it might be used for some malicious purpose by intentionally mimicking some individuals. Since our model is designed to be able to generalize to unseen users, we suggest keeping all personal data (i.e. personal dialogue history) local, as suggested by Dudy et al. (2021), to minimize the risk of malicious imitation. Retrieving from combined history conversations from multiple authors can potentially reduce the risk of exposing personal information of any specific user, but such use is not examined in this work. Finally, we only allow the use of our model on public datasets or under the consent of the individuals being mimicked.

Limitations

In this section, we discuss several limitations of our work that are worth future study.

First, the performance of the hierarchical transformer retriever is limited since the utterance-level transformer is trained from scratch only on our small-scale dataset due to limited time and computational resources. With more resources, future work can focus on pre-training the utterance-level transformer on large-scale data such as the complete Pushshift Reddit data (Baumgartner et al., 2020). Pre-training can potentially improve the performance of the retriever and further improve the generation quality.

Second, the two types of retrieved responses in the RECAP-mixed model are encoded with the same encoder. However, intuitively, the two types of responses should contribute to generation in different ways, so treating them the same way might harm generation performance. This is also reflected in our results. Even though the RECAP-mixed model shows improvement from both types of retrieved responses, the improvement is weaker than that on each separate model. In future work, designing a split encoder for different types of retrieved responses may help maximally preserve the performance boost from both types of retrieved responses.

Acknowledgements

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200006 and in part by DARPA (contracts #HR001121C0169 and #HR00112290025). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. Approved for public release; distribution is unlimited.

References

Jean-Yves Antoine, Jeanne Villaneau, and Anaïs Lefeuvre. 2014. Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 550–559, Gothenburg, Sweden. Association for Computational Linguistics.

JinYeong Bak and Alice Oh. 2019. Variational hierarchical user-based conversation model. 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 1941–1950, Hong Kong, China. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pages 830–839. AAAI Press.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Isabel Briggs-Myers and Peter B. Myers. 1995. Gifts differing: Understanding personality type. *Davies-Black Publishing*.

Robert M Capraro and Mary Margaret Capraro. 2002. Myers-briggs type indicator score reliability across: Studies a meta-analytic reliability generalization study. *Educational and Psychological Measurement*, 62(4):590–602.

Zhangming Chan, Juntao Li, Xiaopeng Yang, Xiuying Chen, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. Modeling personalization in continuous space for response generation via augmented Wasserstein autoencoders. 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 1931–1940, Hong Kong, China. Association for Computational Linguistics.

Hyundong Cho and Jonathan May. 2020. Grounding conversations with improvised dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2398–2413, Online. Association for Computational Linguistics.

Shiran Dudy, Steven Bedrick, and Bonnie Webber. 2021. Refocusing on relevance: Personalization in NLG. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5190–5202, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matej Gjurković, Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Snajder. 2021. PANDORA talks: Personality and demographics on Reddit. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 138–152, Online. Association for Computational Linguistics.

Thamme Gowda, Weiqiu You, Constantine Lignos, and Jonathan May. 2021. Macro-average: Rare types are important too. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1138–1157, Online. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Charles Lee Isbell, Michael Kearns, Satinder Singh, Christian R Shelton, Peter Stone, and Dave Kormann. 2006. Cobot in lambdamoo: An adaptive social statistics agent. *Autonomous Agents and Multi-Agent Systems*, 13:327–354.

Ryota Kishima, Kazuyuki Matsumoto, Minoru Yoshida, and Kenji Kita. 2021. Construction of MBTI personality estimation model considering emotional information. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 262–269, Shanghai, China. Association for Computational Lingustics.

Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. 2020. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4678–4699, Online. Association for Computational Linguistics.

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

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, 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, Barcelona, Spain.

Shilei Liu, Xiaofeng Zhao, Bochao Li, Feiliang Ren, Longhui Zhang, and Shujuan Yin. 2021. A Three-Stage Learning Framework for Low-Resource Knowledge-Grounded Dialogue Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2262–2272, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Xuezhe Ma, Xiang Kong, Sinong Wang, Chunting Zhou, Jonathan May, Hao Ma, and Luke Zettlemoyer. 2021a. Luna: Linear unified nested attention. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 2441–2453.

Zhengyi Ma, Zhicheng Dou, Yutao Zhu, Hanxun Zhong, and Ji-Rong Wen. 2021b. 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.

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

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

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. *CoRR*, abs/1910.10781.

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

David J Pittenger. 2005. Cautionary comments regarding the myers-briggs type indicator. *Consulting Psychology Journal: Practice and Research*, 57(3):210.

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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Yisi Sang, Xiangyang Mou, Mo Yu, Dakuo Wang, Jing Li, and Jeffrey Stanton. 2022. MBTI personality prediction for fictional characters using movie scripts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6715–6724, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2015. Hierarchical neural network generative models for movie dialogues. *CoRR*, abs/1507.04808.

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

Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. 2021. BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. 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 167–177, Online. Association for Computational Linguistics.

Haoyu Song, Wei-Nan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting persona information for diverse generation of conversational responses. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5190–5196. International Joint Conferences on Artificial Intelligence Organization.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chien-Sheng Wu, Andrea Madotto, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020. Getting to know you: User attribute extraction from dialogues. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 581–589, Marseille, France. European Language Resources Association.

Ledell Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes, and Jason Weston. 2018. Starspace: Embed all the things! *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

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

Yujie Xing and Raquel Fernández. 2018. Automatic evaluation of neural personality-based chatbots. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 189–194, Tilburg University, The Netherlands. Association for Computational Linguistics.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069, Florence, Italy. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *CoRR*, abs/1901.09672.

Hanxun Zhong, Zhicheng Dou, Yutao Zhu, Hongjin Qian, and Ji-Rong Wen. 2022. Less is more: Learning to refine dialogue history for personalized dialogue generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5808–5820, Seattle, United States. Association for Computational Linguistics.

A Scientific Artifacts

A.1 Use of Existing Arifacts

Type	Name	License
Dataset	Pushshift Reddit	Not specified
	PANDORA	Not specified
Pre-trained Model	RoBERTa	MIT
	DialoGPT	MIT
	DHAP	Not specified
	MSP	Not specified
	Sentence-RoBERTa	Apache-2.0
	Style-Embedding	MIT
Library	HuggingFace Transformers	Apache-2.0
	Sentence Transformers	Apache-2.0
	Scikit-learn	BSD-3-Clause

Table 6: Licenses of artifacts used in this work.

	# Subsets	Subset Size
Reddit	100	2000
PANDORA		
– Age/Gender	50	100
– MBTI	50	500
Human	20	100

Table 7: T-test hyperparameters.

The licenses of all scientific artifacts used in this paper is shown in Table 6. All artifacts with specified licenses are allowed to use in this work. The PANDORA dataset does not have a license, but we strictly follow their terms of use.[20] All artifacts are intended to be used for research in machine learning and natural language processing, and our use is consistent with this intention.

A.2 Created Arifacts

We release a new model, RECAP in this work under the MIT license. Our model is only intended to be used personally or for research purposes. You can use it for yourself or on publicly available datasets. Using it to mimic other people without authorization is unethical and not allowed.

B T-test Details

For statistical significant tests, we randomly sample subsets from the test set and perform a paired t-test with the subsets' scores. The detailed hyperparameters are shown in Table 7.

C More Experimental Results

The full automatic evaluation results are shown here in Table 8 and Table 9. Metrics within each category in Table 8 are overall consistent with each other. All token-overlap metrics and learning-based metrics are consistent with the human annotated fluency and coherency scores on the top two models. The style metrics are consistent with the human annotated style score on the top two models (with different order for CAV accuracy). Table 9 shows the full personal traits evaluation results. The metrics within each category are less consistent with each other, but the top two models on all metrics are always one of our four models, except for the MBTI J/P score.

[20] https://psy.takelab.fer.hr/datasets/all/pandora/#terms-of-use

Model		Token-overlap Metric				Learning-based Metric		Style Metric	
	PPL↓	BLEU-1↑	BLEU-2↑	ROUGE-L↑	METEOR↑	BERTScore↑	BLEURT↑	Embed Sim↑	CAV Acc↑
DialoGPT	31.25‡	11.54‡	3.26‡	8.49‡	6.86‡	0.4240‡	0.2421‡	22.15‡	51.14‡
DialoGPT + history	29.66‡	12.09‡	3.89‡	9.84‡	7.82‡	0.4361‡	0.2482‡	40.66‡	64.02
DHAP	29.99‡	14.09‡	4.37‡	9.73‡	7.86‡	0.4323‡	0.2511‡	37.14‡	61.53‡
MSP	30.47‡	12.95‡	3.96‡	9.36‡	7.51‡	0.4307‡	0.2453‡	34.72‡	59.61‡
CAP	**29.44**	14.78‡	4.61‡	10.09‡	8.14‡	0.4356‡	0.2534‡	40.38‡	63.71‡
RECAP-style	29.54‡	14.79‡	4.61‡	10.05‡	8.08‡	0.4350‡	0.2525‡	**41.40**	64.14
RECAP-semantic	29.50‡	**15.12**	**4.77**	**10.33**	**8.31**	**0.4617**	**0.2749**	39.65‡	64.00
RECAP-mixed	29.47†	15.06†	4.71†	10.27†	8.27†	0.4372†	0.2557†	40.74†	**64.17**

Table 8: The automatic evaluation results on the Reddit dataset with perplexity, token-overlap metrics, learning-based metrics, and style metrics. The best and second best results in each column are shown in **bold** and underline. "†" and "‡" indicates statistically significant difference for $p < 0.05$, between the best or the top two models, respectively, determined by t-test.

Model	Demographic		MBTI				
	Age↑	Gender↑	I/E↑	S/N↑	T/F↑	J/P↑	Average↑
DialoGPT	0.0425‡	0.6103‡	0.5013‡	0.4982‡	0.5317‡	0.4655‡	0.4992‡
DialoGPT + history	0.1008‡	0.6763‡	0.4986‡	**0.5146**	0.5561‡	0.4826‡	0.5130‡
DHAP	0.0829‡	0.6653‡	0.4997‡	0.4872‡	0.5473‡	0.5135†	0.5119‡
MSP	0.1226‡	0.6816‡	0.4832‡	0.4924‡	0.5449‡	0.4870‡	0.5019‡
CAP	**0.2051**	0.7077†	0.4998‡	0.4965‡	0.5635†	0.5070†	0.5167†
RECAP-style	0.1822	0.7001†	0.5167†	**0.5146**	0.5562‡	**0.5185**	0.5265
RECAP-semantic	0.1699†	**0.7303**	**0.5279**	0.5124	0.5676	0.5025‡	**0.5276**
RECAP-mixed	0.1392‡	0.6962†	0.5154†	0.5045‡	**0.5719**	0.5049‡	0.5242
Ground-truth	0.2617	0.7477	0.5149	0.5064	0.5705	0.5108	0.5257

Table 9: The automatic personal traits evaluation results on the PANDORA dataset. The best and second best results in each column are shown in **bold** and underline. "†" and "‡" indicates statistically significant difference for $p < 0.05$, between the best or the top two models, respectively, determined by t-test.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 8 Limitations.

☑ A2. Did you discuss any potential risks of your work?
Section 7 Ethical Issues.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1 Introduction.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Appendix B Scientific Artifacts summarizes the scientific artifacts we used and created in this work.

☑ B1. Did you cite the creators of artifacts you used?
Along with the first occurrence of each artifact in our paper.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix B Scientific Artifacts

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix B Scientific Artifacts

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We did not collect data in this work. We used two widely used existing datasets and refer reviewers to the works describing these datasets for details on data collection issues (https://arxiv.org/abs/2001.08435, https://arxiv.org/pdf/2004.04460v3.pdf) . We did not perform any further anonymization or offensive content filtering to preserve persona and personal writing style as much as possible.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Along with the first occurrence of each artifact in our paper, we add a footnote for the documentation link (or github link if documentation is not provided).

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3.1 Dataset and Section 3.4.1 Automatic Evaluation Personal Traits Metrics

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Section 3 Experiments & Appendix A Computational Experiments Details

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A Computational Experiments Details

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3.3 Implementation Details and Appendix A Computational Experiments Details

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4 Results

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3.3 Implementation Details and Appendix A Computational Experiments Details

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 3.4.2 Manual Evaluation & Section 4.2 Human Evaluation Results

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 3.4.2 Manual Evaluation & Section 4.2 Human Evaluation Results

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 3.4.2 Manual Evaluation. We did not discuss payment since the annotators are volunteers.

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
We do not collect any data in this work. We only use annotators for results evaluation. See B.4.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The conversational data was previously collected and published by third parties. See B.4.; We submitted a protocol for human annotations of natural language to our IRB, who determined this to not be HSR, and as such was exempt from further review.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We do not collect any data in this work; see B.4. We only use annotators for results evaluation. We describe the two (internal) annotators used.