

Article

Persona-Identified Chatbot through Small-Scale Modeling and Data Transformation

Bitna Keum ¹, Juoh Sun ¹, Woojin Lee ¹, Seongheum Park ¹ and Harksoo Kim ^{2,*}

¹ Department of Artificial Intelligence, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Republic of Korea; beausty23@konkuk.ac.kr (B.K.); qssz1326@konkuk.ac.kr (J.S.); shes100@konkuk.ac.kr (W.L.); tjdgma95@konkuk.ac.kr (S.P.)

² Division of Computer Science and Engineering & Department of Artificial Intelligence, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Republic of Korea

* Correspondence: nlpdrkim@konkuk.ac.kr; Tel.: +82-2-450-3499

Abstract: Research on chatbots aimed at facilitating more natural and engaging conversations is actively underway. With the growing recognition of the significance of personas in this context, persona-based conversational research is gaining prominence. Despite the abundance of publicly available chit-chat datasets, persona-based chat datasets remain scarce, primarily due to the higher associated costs. Consequently, we propose a methodology for transforming extensive chit-chat datasets into persona-based chat datasets. Simultaneously, we propose a model adept at effectively incorporating personas into responses, even with a constrained number of parameters. This model can discern the most relevant information from persona memory without resorting to a retrieval model. Furthermore, it makes decisions regarding whether to reference the memory, thereby enhancing the interpretability of the model's judgments. Our CC2PC framework demonstrates superior performance in both automatic and LLM evaluations when compared to high-cost persona-based chat dataset. Additionally, experimental results on the proposed model indicate the improved persona-based response capabilities.

Keywords: persona-based dialogue; long-term conversation; persona memory; low cost; data transformation



Citation: Keum, B.; Sun, J.; Lee, W.; Park, S.; Kim, H. Persona-Identified Chatbot through Small-Scale Modeling and Data Transformation. *Electronics* **2024**, *13*, 1409. <https://doi.org/10.3390/electronics13081409>

Academic Editor: Flavio Canavero, Bo Xu, Linmei Hu and Jian Liu

Received: 15 January 2024

Revised: 6 March 2024

Accepted: 4 April 2024

Published: 9 April 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With significant advancements in neural models, open-domain dialogue systems have seen substantial improvements. Research is being actively conducted to develop more natural and human-like chatbot. In relation to this, personas are crucial, which encompass key information about the speaker such as personal information, preferences, and values [1,2]. In human conversations, humans memorize meaningful information about the counterpart. In subsequent conversations, they respond by leveraging this memorized information. By incorporating personas into conversations, this approach enables a closer mimicry of human conversational abilities. Therefore, various datasets for persona-based conversations have been developed. These datasets are predominantly created through human annotators [3–6], which incurs significant costs. From online communities, it is possible to inexpensively obtain a substantial volume of data [7,8], but the quality of this data may be notably poor. Additionally, to closely mimic real-world conversational scenarios, datasets consisting of multiple dialogue sessions have been proposed [9–11]. The term “session” denotes each dialogue that is split based on brief temporal pauses. These datasets have demonstrated exceptional efficacy in fostering long-term conversational capabilities. However, their construction demands significantly more extensive tasks from human annotators. These annotators are provided with a memory that stores personas acquired from previous sessions, and are instructed to engage in conversations while

referencing them. However, there are significant challenges in this construction process that make it prohibitively expensive.

In light of this constraint, we propose the CC2PC (Chit-Chat to Persona-Chat) framework, which transforms abundant chit-chat data resources into persona-based chat. We derive insights from the construction process of Multi-Session Chat (MSC) [9], which has proven effectiveness in long-term conversations. Consequently, targeting single-session dialogues without predefined personas, we have developed a method to create a self-derived memory by feeding back the summaries generated from the current conversation as inputs to the same conversation. As a result, since the memory contains summaries of the current dialogue, it enables the transformation of conventional dialogue into persona-based conversations that reference the provided memory. Through this automated process, significant cost savings can be achieved, along with the acquisition of a substantial volume of data.

To enhance the capabilities of persona-based chat, various dialogue models have been investigated. While the majority of research focuses exclusively on the model's effectiveness in reflecting the persona within its responses, it is important to recognize that humans do not consistently utilize their memories at every turn. Responses are often made by the ongoing context of the conversation. Therefore, while the capability to reflect memory well is undeniably significant, the prerequisite skill lies in discerning whether a memory reference is required before actualizing the act of memory reference.

Most prior studies utilized a separate retriever to extract pertinent element from the memory. However, within the retrieval-augmented generation (RAG) [12] framework, the dialogue model can be fatally influenced by the performance of the retrieval model. Furthermore, there may be notable slowdown and diminished efficiency in the speed of training and inference steps. Therefore, we propose a model that not only discerns the necessity of referencing memory but also identifies pertinent memory elements and integrates them into the response. By employing a retrieval-free methodology, internal computations within the generator are executed to enhance system efficiency. Notably, attention computations occur at two distinct levels between the dialogue context and persona memory, facilitating a comprehensive understanding of the relationship from multiple perspectives. The model's ability to discern the persona to be referenced is improved through supervised learning with the use of pseudo-labels. The architecture branching that is based on the prediction of the necessity of memory reference leads to a more streamlined and efficient structure.

Diverse experimental results demonstrate the superior performance of our CC2PC framework, surpassing high-cost crowdsourced datasets. Notably, the performance gap becomes more pronounced in later sessions, suggesting that our framework can effectively construct data aligned with long-term conversational environments similar to the real world.

The main contributions of this study are as follows:

- (1) We introduce a novel framework that transforms abundant chit-chat data resources into persona-based chat. This enables the acquisition of a large volume of persona-based chat data at minimal cost.
- (2) We propose an efficient model designed to identify the pertinent memory elements and determine whether to reference the memory, using a retrieval-free approach.
- (3) The experiment results demonstrate that the low-cost data obtained through our framework surpasses the performance of high-cost, human-annotated datasets, especially in long-term conversations. This indicates that our constructed data is well-aligned with the requirements of long-term conversational contexts.

2. Related Work

2.1. Persona-Based Dialogue Datasets

In the field of persona-based dialogue systems, PERSONA-CHAT [3] is the initial persona-based dialogue dataset constructed by crowdworkers. They collected data by assigning random personas to paired workers, who were then instructed to engage in

dialogue based on these personas. Inspired by this approach, the DuLeMon dataset [6] was built with the additional process of labeling grounded personas. The Knowledge Behind Persona (KBP) dataset [13] is a dataset for personalized knowledge-grounded dialogues. It uses a portion of the DuLeMon dataset as a seed and matches persona and knowledge using Chinese knowledge bases. Then, human annotators create dialogues and label grounding information. While intuitive and effective, crowdsourcing-based approaches entail substantial costs. In an effort to reduce construction costs, several datasets that are collected from online communities have also been proposed [7,8]. They collected data by pairing posts and comments and considering each pair as one dialogue turn. Owing to abundance of data available on the web, this approach offers the advantage of being able to construct a vast amount of data at a low cost. However, most of these datasets consist of single-turn dialogues with relatively short lengths of utterances. Furthermore, there is a potential for low data quality and privacy concerns because of the nature of web-crawled data. Meanwhile, as existing datasets are primarily composed of single dialogues, they may not effectively support research on long-term conversations that represent real-world scenarios. To overcome this limitation, several datasets have been developed expanding from a single session to encompass multiple sessions. Xu et al. [9] proposed the MSC dataset for long-term conversations. Built upon PERSONA-CHAT [3] dataset, they extended the dialogues from a single session to multiple sessions via crowdsourcing. Instead of the full context from the previous session—which could prolong comprehension time and lead to inefficiency—summaries from personal traits perspective on the context were stored in the memory. In the subsequent session, the conversation proceeded based on this memory. The overall processes were repeated over multiple sessions. Bae et al. [10] built a dataset by using crowdsourcing as the MSC approach, but also incorporating a large language model (LLM) to reduce the collection costs. However, these approaches for constructing multi-session dialogue datasets have a critical limitation—exorbitant costs due to the more sophisticated tasks compared to basic crowdsourcing.

Therefore, to address this gap, there is a crucial need for a methodology that constructs persona-based dialogue data in a cost-effective manner and ensures effectiveness even in long-term conversations. Through our CC2PC framework, it becomes possible to realize both aspects. The construction methods of datasets proposed previously are predominantly manual or semi-automatic, leading to significant time and cost expenditures. In contrast, our CC2PC framework employs a fully automatic approach, offering cost efficiency. Despite the automation, it maintains high dialogue quality by utilizing conversations as they were originally written by humans. Moreover, our framework reflects key properties of the dataset constructed for long-term conversations, which is why it is effective even in long-term conversations. Meanwhile, existing persona-based dialogue datasets lack labels indicating which persona was referenced in responses. Accordingly, it was hard to explicitly train the model to distinguish between cases where responses should be based on personas and cases where responses should be based on the dialogue context. However, datasets created using the CC2PC framework are labeled with referenced personas, enabling training on this aspect.

2.2. Persona-Based Dialogue Model

Research on generative models are increasingly focused on building chatbots that exhibit human-like consistency in personality [1,2]. The first persona-based neural model [1], known as the Speaker model, used speaker embeddings to generate speaker-specific responses. Following this, Zhang et al. [3] implemented a Seq2seq model with attention, encoding personas into memory for reference during response generation. Building on this concept, Wu et al. [14] divided the memory into separate components for profile and comment history. TransferTransfo [15], based on the transformer architecture, incorporates persona information into input embeddings and employs transfer learning, extending its capabilities to multi-task learning for both generation and classification tasks. The Persona Perception Bot (P^2 Bot) [16] supports interaction between the transmitter and receiver, enabling the

interlocutors to perceive the mutual persona and use this information to generate responses. This model leverages both supervised learning (SL) and reinforcement learning (RL) [17] in training phase. CharacterGLM [18] is a large-size model for character-based dialogues. This model allows for the customization of AI characters or social agents based on user-defined attributes and behaviors. The attributes represent static or gradual features of the character, while behaviors encompass dynamic elements such as linguistic characteristics, emotional expressions, and interaction patterns. Zhou et al. [19] introduces SUPPORTER, a model that uses reinforcement learning for emotional support conversations. The model balances empathy and elicitation while maintaining coherence. It employs a unique reward system that adjusts elicitation intensity, considers user feedback, and ensures dialogue coherence. Kwak et al. [20] proposes a model that generates context-dependent instructions for each dialogue turn, thereby enhancing the quality of responses. By aligning instructions to the input during instruction tuning, the model adapts to the intricate conditions of each dialogue. It underscores the importance of personalized guidance in multi-turn conversations. The SAFARI framework [13] uses LLMs to manage multiple knowledge sources in dialogue systems. The process is divided into planning, retrieval, and assembling stages. In planning, LLMs decide on the usage and order of knowledge sources. Relevant knowledge is then retrieved, and the final response is generated by incorporating this knowledge and the dialogue context.

Meanwhile, there are various models based on the RAG framework, which utilizes a separate retrieval model to search relevant persona from the memory and conveys the result to the generation model [6,9,21]. However, having a separate retrieval model makes the entire system more burdensome and slows down the inference speed. Therefore, we propose a model that can autonomously identify relevant information in the memory without using a separate retrieval model. In addition, the proposed model can provide two types of responses: based solely on the dialogue or based on the persona. This is accomplished by identifying the presence of persona relevant to the dialogue context.

3. Materials and Methods

3.1. CC2PC Framework

Xu et al. [9] demonstrated that building dialogues into multiple sessions is effectual in long-term conversations. However, this approach incurs enormous costs. Therefore, despite the lack of data reflecting long-term conversational situations, it is challenging to construct it. Taking this into consideration, we propose a budget-friendly data transformation method based on insights gained from the data collection method of MSC. In MSC, memory containing summaries of the previous session is provided on the current session. The utterances in the current session reference the memory of the previous session. In our circumstance, however, we do not have a previous session because we will use single session dialogues. Therefore, instead of creating utterances that reference the memory of the previous session, we take an inverse approach by creating memory treated as referenced by the utterances in the current dialogue. For this approach, the first step is summarizing important information from the current dialogue in terms of personal traits. After additional processes, the persona summary is stored in the memory as a ground-truth persona referenced by the utterance in current dialogue. The reason for using it as a ground-truth persona is that an utterance (e.g. "Oh, pasta is my favorite food!") aligns content-wise with its summary (e.g. "I love pasta."), making the persona derived from the summary a suitable reference for that utterance.

In a nutshell, this framework transforms existing chit-chat dialogues into persona-based dialogues that reference the ground-truth persona in the given memory. We similarly simulated the collection method of MSC by using only single-session dialogue datasets. The following subsections provide detailed descriptions of the data construction process depicted in Figure 1. The data statistics of the transformed training set are shown in Table 1.

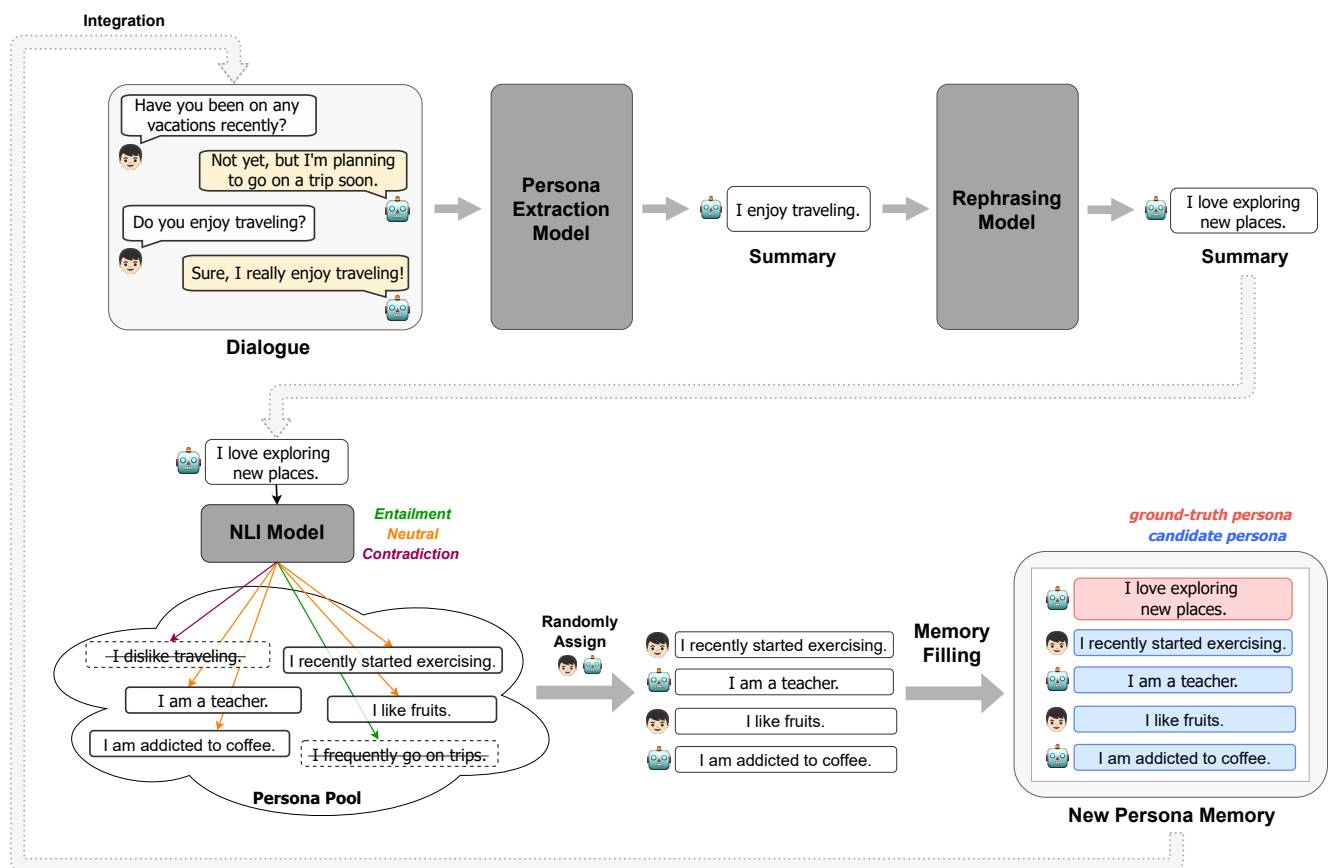


Figure 1. Flowchart of the CC2PC framework. Summaries obtained through persona extraction and rephrasing in existing dialogues serve as ground-truth personas. Using an NLI model, neutral personas related to the ground-truth persona are selected as candidate personas. Ground-truth and candidate personas are then filled into the memory. Finally, this new persona memory is integrated into the initial dialogue.

Table 1. Data statistics of the transformed training set. The symbol ‘#’ indicates number.

Statistics	Number
# of Dialogues	10,000
# of Utterances	56,160
Avg. # of utterances	5.6
Avg. length of utterances	21.5
Avg. # of personas	17.1
Avg. # of bot personas	10.9
Avg. # of user personas	6.2

3.1.1. Persona Extraction

Unlike persona-based chat datasets, chit-chat datasets do not include personas. Therefore, we aim to use summaries of the current session’s dialogues as personas. We utilize a persona extraction model (<https://parl.ai/docs/zoo.html#dialogue-summarization-model> (accessed on 15 January 2024)). It extracts and generates persona summaries from the dialogues. To construct a target utterance with only utterances from the bot, we excluded summaries derived from the user’s utterances from the generated personas.

3.1.2. Persona Rephrasing

If we construct training data by using summaries of the target utterances directly as persona, the chatbot model trained with this data may potentially face issues with lexical diversity, as there is a possibility that it could generate utterances by simply replicating the

tokens of the persona. Indeed, measuring the lexical overlap between the target utterances and their summaries using the BLEU metric revealed overlap scores of 15.94 in DailyDialog (DD) [22], 18.58 in Empathetic Dialogues (ED) [23], and 11.41 in Topical-Chat (TC) [24]. This suggests that if summaries are used as they are as persona memory, the response generation may be trained to merely replicate the persona's vocabulary. Therefore, to enable the generation of diverse vocabularies based on the persona, we performed rephrasing on the summaries using a Rephrasing Model. This rephrasing model was composed of a T5 model [25] fine-tuned with the chatgpt-paraphrases data (<https://huggingface.co/datasets/humarin/chatgpt-paraphrases> (accessed on 15 January 2024)). The results showed that the lexical overlap, as measured by the BLEU metric, decreased to specific scores for each dataset: from 15.94 to 6.02 for DD, from 18.58 to 6.20 for ED, and from 11.41 to 4.48 for TC. Concurrently, in the DD dataset, the BERTScore changed minimally from 0.90 to 0.88, in the ED dataset from 0.92 to 0.90, and in the TC dataset from 0.90 to 0.88, maintaining semantic similarity.

3.1.3. Persona Memory Filling

If using all the summaries from the current session, it can be hard to decide which one to reference, as they often have similar context. Therefore, we constituted the memory only with the personas labeled on the target utterance, which are called as ground-truth personas. However, in general persona-based dialogue datasets, there also exist personas that are not relevant to the target utterance. Therefore, the next step is structuring the memory to include not only the ground-truth persona but also irrelevant personas. Beforehand, we constructed a persona pool consisting of 30,000 personas, directly sourced from the PERSONA-CHAT training set. We randomly sampled some personas from this pool. We then infer the relation between the ground-truth persona and the sampled persona through natural language inference (NLI). As the NLI model, we employed the RoBERTa-large [26] model, which had been fine-tuned on both the MNLI [27] and DNLI [28] datasets. Personas inferred as Entailment or Contradiction were filtered out. Among the personas inferred as Neutral, only a maximum of five are selected and assigned a random speaker (user or bot), which are called as candidate personas. Ground-truth and candidate personas are then filled into the memory. Finally, this new memory is integrated into the initial dialogue, structuring the data as if the persona memory had been given earlier than the dialogue.

3.2. Proposed Model

3.2.1. Task Definition

Figure 2 illustrates the architecture of the proposed model. The objective of our model is to generate a response R , given a persona memory $M = \{p_1, p_2, \dots, p_m\}$ and a dialogue $D = \{u_1, u_2, \dots, u_n\}$. Here, M consists of m persona sentences, and D consists of n utterances exchanged alternately between two speakers. First, we employ two separate encoders (i.e., persona encoder and dialogue encoder) to encode D and each sentence in M . Based on the embeddings of two encoders, our model calculates Sentence-Level Attention (SLA) and identifies whether there is a relevant persona within M . Based on the results of Persona Identification, the Architecture branching is executed. If the relevant persona does not exist, it is considered that no memory reference is needed, and thus, the dialogue embedding is passed to the decoder. Otherwise, it is considered that memory reference is needed, and thus, our model further computes Token-Level Attention (TLA) between the two embeddings. It then combines the outputs from SLA and TLA to consider both semantic and lexical information. Subsequently, it feeds the combined representation into the decoder to generate a response incorporating the relevant persona. We elaborate on our model in the remainder of this section.

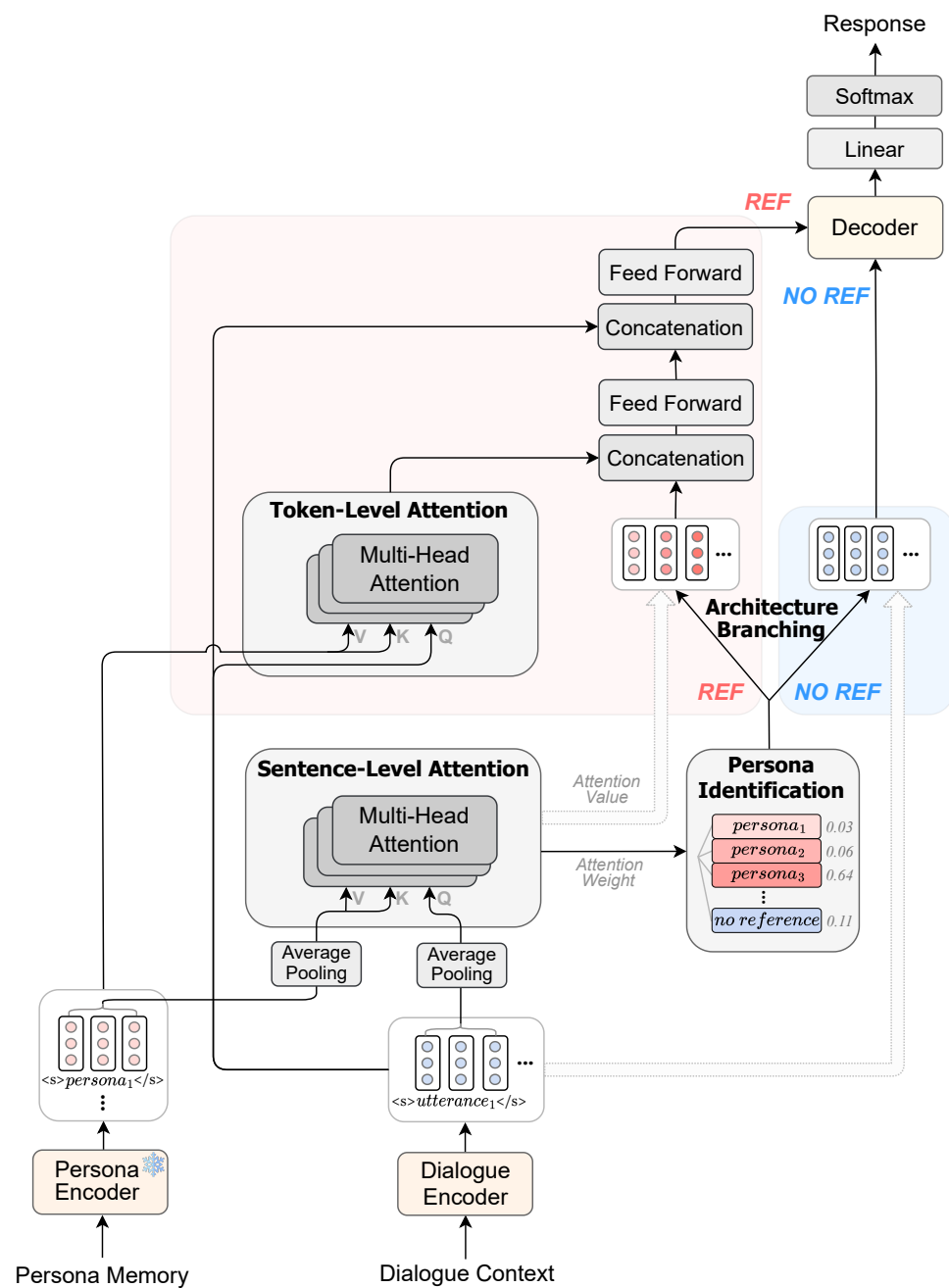


Figure 2. The overall architecture of the proposed model. It consists of the bi-encoder structure with persona encoder and dialogue encoder. It comprises four components: Persona Identification, Sentence-Level Attention, Token-Level Attention, and Architecture Branching.

3.2.2. Encoder

We adopt the bi-encoder structure [29] that uses separate encoders for the persona and the dialogue context. Therefore, it is possible to encode instances of memory in advance, resulting in faster inference speed. Both of the encoders are initialized with the pretrained parameters of BART [30].

Dialogue Encoder

The dialogue encoder takes a dialogue sequence D as input, where an EOS token (i.e., $\langle /s \rangle$) is inserted between each utterance and the tokens of all utterances are concatenated. The dialogue encoder then outputs contextualized token vectors for each token

in D . Let an utterance $u_i = \{t_1^{u_i}, t_2^{u_i}, \dots, t_{c_i}^{u_i}\}$, where T_i denotes the number of tokens in the i -th utterance. Then, the dialogue encoder generates the contextualized embedding $H^D = \{h_1^{u_1}, h_2^{u_1}, \dots, h_{c_1}^{u_1}, \dots, h_1^{u_n}, h_2^{u_n}, \dots, h_{c_n}^{u_n}\}$. It is possible to finely understand the relationships among all tokens within the context through the self-attention operation of the encoder.

Persona Encoder

The personas in memory M are all independent; therefore, each persona is encoded separately, as there is no need to understand the relationships among personas. Given the memory M , the persona encoder generates contextualized token vectors for each token within p_i . Let a persona $p_i = \{t_1^{p_i}, t_2^{p_i}, \dots, t_{k_i}^{p_i}\}$, where k_i denotes the number of tokens in the i -th persona sentence. All of the encoded outputs of each persona are then concatenated, forming the memory embedding $H^M = \{h_1^{p_1}, h_2^{p_1}, \dots, h_{k_1}^{p_1}, \dots, h_1^{p_m}, h_2^{p_m}, \dots, h_{k_m}^{p_m}\}$. Unlike the dialogue encoder, the persona encoder remains frozen during training.

3.2.3. Persona Identification

The objective of this module is to improve model's abilities of determining the necessity of memory reference and identifying the most relevant persona. First, average pooling is executed on token vectors within each sentence to obtain sentence vectors, which represent the sentence's global meaning. Using these, SLA is computed to calculate the semantic similarities between the dialogue and personas. The weights of SLA, denoted as $W_{SLA} \in \mathbb{R}^{n \times m}$, play a crucial role in comprehending semantic relevance.

$$W_{SLA} = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \quad (1)$$

$$= \begin{bmatrix} \hat{y}_{1,1} & \dots & \hat{y}_{1,m} \\ \vdots & \ddots & \vdots \\ \hat{y}_{n,1} & \dots & \hat{y}_{n,m} \end{bmatrix}$$

Based on W_{SLA} , the instance of M with the highest weight is used as the prediction. If a ground-truth persona exists, it is used as the answer label; otherwise, NO_REFERENCE is used as the answer label, which indicates that memory reference is unnecessary when there is no relevant persona in memory. With the prediction and the label, we conduct supervised learning. For the loss function, we use the cross-entropy loss:

$$\mathcal{L}_{PI} = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m y_{i,j} \log(\hat{y}_{i,j}) \quad (2)$$

The $\hat{y}_{i,j}$ is the model's predicted probability that the i -th utterance should reference the j -th persona, and $y_{i,j}$ is the answer label.

3.2.4. Architecture Branching

In this section, we explain how the architecture branches based on the need for memory reference.

If the persona identification module determines that memory referencing is required, TLA is further computed. This step is crucial for conducting in-depth analysis of how the dialogue is lexically associated with the personas. The final representation is obtained by integrating the outputs of SLA and TLA and combining the dialogue embedding to it. It is fed into the decoder.

Otherwise, in situations where memory referencing is not required, the model simplifies the process by skipping SLA and TLA and directly feeding the dialogue embedding into the decoder. This approach allows the model to focus solely on the dialogue. This reduces unnecessary computational load and enhances the system's efficiency.

3.2.5. Training Objective

We compute the final training loss \mathcal{L} as follows:

$$\mathcal{L} = \lambda \mathcal{L}_{RG} + (1 - \lambda) \mathcal{L}_{PI} \quad (3)$$

The first term \mathcal{L}_{RG} represents the loss pertaining to response generation, and the second term \mathcal{L}_{PI} represents the loss for persona identification. The weight of each loss component is determined by the hyper-parameter λ . In our experimental setup, we assigned the value of λ to 0.8.

4. Experiments

4.1. Datasets

In our study, we focused on the utilization of three significant datasets: DailyDialog (DD) [22], Empathetic Dialogues (ED) [23], and Topical-Chat (TC) [24]. The DD dataset comprises 8300 dialogues, ED includes 11,000 dialogues, and TC comprises 47,000 dialogues. These datasets were specifically chosen for their diversity and depth in conversational contexts.

For our evaluation, we used the MSC test dataset, which exhibits statistics as shown in Table 2. In session 1, with an amount of data compared to other sessions, we observe the lowest percentage of instances requiring references to persona memory. This implies a need for judgment in identifying the situation when a memory reference is necessary. As the sessions progress, there is an increase in the number of persona memories, and we also observe a rise in the proportion of conversational data that requires memory referencing in the test data. This indicates that with advancing sessions, not only the ability to judgment whether to reference or not becomes more critical, but also the skill to determine which specific memory to reference increases in importance.

Table 2. Data statistics of MSC test set. S1, S2, S3, S4, and S5 denotes the session 1, 2, 3, 4 and 5, respectively. The symbol ‘#’ indicates number.

Statistics	S1	S2	S3	S4	S5
# of Dialogues	966	500	500	500	500
# of Utterances	14,041	5400	5376	5397	5398
Avg. # of utterances	14.5	10.8	10.7	10.8	10.8
Avg. length of utterances	21.5	13.4	29.1	32.5	34.2
Max # of personas	5	12	16	22	27
Avg. # of personas	4.5	7.8	10.4	14.1	17.3
Avg. # of bot personas	4.5	3.9	5.3	7.2	8.8
Avg. # of user personas	0.0	3.8	5.1	6.9	8.5
# of pseudo-labeled personas	231	223	337	370	393
Persona referencing ratio	0.2	0.4	0.7	0.7	0.8

4.2. Baselines

we select the BART_{large} [30] as the baseline. BART stands for Bidirectional and Auto-Regressive Transformers, which is trained via a denoising approach using extensive text data. This model is recognized as a powerful pre-trained language model, exhibiting exceptional performance particularly in text generation tasks. It comprises 12 Transformer Encoder layers and 12 Transformer Decoder layers. Each layer features a hidden size of 1024 and 16 attention heads, accumulating to a total of 406 million parameters. This model is accessible through the Transformers library of HuggingFace (<https://huggingface.co/facebook/bart-large> (accessed on 15 January 2024)).

We also use ChatGPT and Llama-2 as comparison models. ChatGPT is a representative closed-source LLM, and Llama-2 is a representative open-source LLM. For ChatGPT, we used the gpt-3.5-turbo-1106 (<https://platform.openai.com/docs/models/gpt-3-5-turbo> (accessed on 15 January 2024)) API provided by OpenAI to perform experiments. Because ChatGPT is a commercial black box product, there is no officially disclosed information about its parameter size. For Llama-2, we use Llama-2 7B model served by ollama

(<https://ollama.com/library/llama2> (accessed on 15 January 2024)) framework to perform comparison experiments.

4.3. Experimental Settings

We use the batch size of 8 and apply gradient accumulation steps of 4. The learning rate is set to 3×10^{-5} with AdamW as the optimizer. We trained the model for 10 epochs. In order to generate responses, we use beam search as a decoding method with beam size 8, and the repetition penalty as 1.1.

4.4. Quantitative Evaluation

To evaluate the effects of the CC2PC framework and the proposed model, we used metrics such as Perplexity (PPL), BLEU, Distinct-2 (DIST-2), and BERTScore (BS) for quantitative evaluation.

Perplexity

PPL is a metric used to measure the uncertainty of a language model in predicting a sequence of words. A lower PPL indicates that the model is more accurate in its predictions, reflecting better understanding and fluency in the language. The formula for this is shown below. $X = \{x_1, x_2, \dots, x_t\}$ represents the input sequence, and t denotes the total number of tokens in the input sequence.

$$\text{Perplexity}(X) = \exp\left(-\frac{1}{t} \sum_{i=1}^t \log p_{\theta}(x_i | x_{<i})\right) \quad (4)$$

BLEU

The BLEU evaluates the linguistic similarity between the generated text and the reference text by calculating the number of matching n-grams, which are contiguous sequences of words. A higher BLEU score indicates greater similarity, demonstrating that the generated text closely aligns with the language and structure of the reference text. The formula for the BLEU is presented below. w_n is the weight for each n-gram precision, and p_n is the n-gram precision. r is the length of the reference sentence, and c is the length of the generated sentence. The Brevity Penalty (BP) is a factor that reduces the score when the length of the generated sentence is shorter than the length of the reference sentence, serving as a correction mechanism to ensure that the generated sentence is not too short.

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^N w_n \log(p_n)\right) \quad (5)$$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{else} \end{cases}$$

Distinct-2

DIST-2 evaluates the diversity of the generated text by analyzing the proportion of unique bi-grams in the text. A higher DIST-2 score indicates a more diverse vocabulary, suggesting that the text contains a wider range of word pairs, thus reflecting the model's ability to generate varied and rich language. The formula is presented below.

$$\text{Distinct-}n = \frac{\text{Number of distinct } n\text{-gram tokens}}{\text{Total number of generated tokens}} \quad (6)$$

BERTScore

BS evaluates the semantic similarity by using BERT embeddings to calculate the cosine similarity, assessing how closely the meaning of the generated text aligns with that of a reference text. The formula is presented below. $\text{Precision}_{\text{BERT}}$ is calculated by finding the token in the reference sentence that has the highest cosine similarity for each token in the candidate sentence, and then taking the average of these similarity values. $\text{Recall}_{\text{BERT}}$

is calculated by finding the token in the candidate sentence that has the highest cosine similarity for each token in the reference sentence, and then taking the average of these similarity values.

$$\text{BERTScore} = \frac{2 \times \text{Precision}_{\text{BERT}} \times \text{Recall}_{\text{BERT}}}{\text{Precision}_{\text{BERT}} + \text{Recall}_{\text{BERT}}} \quad (7)$$

4.5. LLM-Based Evaluation

Numerous investigations have demonstrated that Large Language Models (LLMs), primarily ChatGPT and GPT-4, surpass human capabilities in various tasks [31–33]. Therefore, we selected the advanced GPT-4 model as the LLM evaluator. We used the GPT-4 Turbo through the API provided by OpenAI. We had given the LLM evaluator a prompt to evaluate responses. The prompt used for evaluation is presented in Figure 3. We randomly sampled data for evaluation from the MSC test set. To avoid contamination by other data samples, we conducted the evaluation in independent session for each sample. The code used for evaluation can be found in Appendix B.

For the evaluation metrics, we use Consistency, Fluency, and Engagingness. Scores for the metrics range from 0 to 2.

Consistency

Consistency involves evaluating whether each response logically follows the preceding dialogue and accurately reflects the characteristics and behaviors outlined in the provided personas.

Fluency

Fluency evaluates the grammatical correctness and the presence of awkward or repetitive representations in the responses.

Engagingness

Engagingness denotes the level of interest in a response, encompassing aspects of informativeness and attractiveness.

Read the [Personas] of each speaker and the [Conversation] carefully. Please evaluate the [Response 1] and [Response 2] based on them. For each response, assign a score of 0, 1, or 2 for the following metrics:

- 1. Consistency:** Assess whether the response aligns naturally with the ongoing conversation and remains consistent with the provided personas, without any contradictions.
- 2. Fluency:** Evaluate the response for its grammatical correctness and absence of awkward repetitions.
- 3. Engagingness:** Determine if the response is interesting and captivates the attention of the audience.

Please evaluate each Response independently. Provide only the scores for each metric without detailed explanations.

[Personas]

partner's persona: I teach jiu-jitsu to marines.
partner's persona: I plan on becoming an allergist. I love milk, and I hope to drink more of it in the future.
your persona: I prefer scuba diving over martial arts. I love snakes.
partner's persona: I am currently in college. I will be taking summer courses at my college.
your persona: I am on a vacation in Hawaii, which is my third time doing so. I do not like school.
your persona: I got divorced recently.
your persona: I was a marine in the Vietnam war. I ate lizards back then.
partner's persona: I am allergic to snakes and dairy.

[Conversation]

partner: I was talking with my boss a while ago regarding the martial arts class I was teaching, and they mentioned the class will continue in the summer. If your interested, I can help you enroll so you can learn jiu-jitsu this summer.
you: Omg that would be perfect thank you so much. Just let me know the details of what you need from me and i will do it ASAP.
partner: Absolutely! Maybe you could get me into scuba diving, too!
you: Yes of course. That would be such great fun. The next time I go I will for sure invite you. I think you will love it.
partner: Where's the best place in the world for scuba diving, in your opinion?
you: Honestly my all time favorite place that i have done Scuba Diving has to be in Bora Bora. It was a beautiful experience.
partner: Oh wow! Did you fly there, or did you sail out there? A friend of mine once did Tahiti and Bora Bora on a yachting adventure!
you: I flew there. But oh my next time i defiantly have to do a yachting adventure that must of been quite the experience.
partner: It sounded like an incredible time. One day when I win the lottery, I'm there lol.
you: Yeah it must be quite an expensive adventure, but one i for sure need to look into.
partner: Are you retired now? I'm guessing you're living a pretty comfortable life these days!
you:

The followings are response candidates of you in the last turn of the [Conversation].

[Response 1]: *i'm retired now, but i used to be a real estate agent.*
[Response 2]: *I'm retired, but still living at home with my wife and two kids. We're just starting to plan our honeymoon to Tahitian getaway!*

Figure 3. Prompt example used for LLM-based evaluation. Bold text represents the fixed template, while italicized text represents variable text.

4.6. Qualitative Evaluation

We also conducted qualitative assessments to evaluate the performance of the chatbot, a traditional method where the evaluation is directly carried out by humans. Despite its high time and cost demands, this method allows for a precise evaluation that reflects human preferences. We utilized the same dataset that was randomly sampled for the LLM-based evaluation. The evaluation was carried out by three researchers who were not involved in the model development but are well-versed in dialogue systems. The evaluation metrics, consistent with those used in the LLM-based evaluation, included Consistency, Fluency, and Engagingness, with scores ranging from 0 to 2.

5. Results

Examples of dialogue responses generated by the model can be found in Appendix A.

5.1. Data-Related Experiments

5.1.1. Experiment on the Dataset Combination

We conducted an experiment to evaluate performance based on combinations of various datasets. We tested combinations of DD + ED, ED + TC, and TC + DD, alongside an integrative approach that combined all three datasets (DD + ED + TC). The total number of data in each combination was equal, and the number of data used in each combination was evenly divided.

The results of the experiment are presented in Table 3. In the analysis of results from session 1, notable findings were observed. The DD + ED combination resulted in PPL of 46.21 and a BLEU score of 0.08, while the ED + TC combination showed significant improvements with a PPL of 20.24 and a BLEU score of 0.71. Similarly, the TC + DD combination yielded a PPL of 22.61 and a BLEU score of 0.53. However, the most striking results emerged from the combined DD + ED + TC dataset. This comprehensive combination demonstrated the best performance across all evaluated metrics, with a notably lower PPL of 17.68 and a higher BLEU score of 0.56. This trend was similarly observed in other sessions.

Based on these findings, we concluded that employing the combination of all three datasets—DD, ED, and TC—is the most effective strategy.

Table 3. Quantitative evaluation results for data combinations. DD denotes the DailyDialog dataset, ED denotes the Empathetic Dialogues dataset, and TC denotes the Topical-Chat dataset. Bold indicates the best performance across dataset combinations for a specific metric in each session.

Session	Dataset Combination															
	DD + ED				ED + TC				TC + DD				DD + ED + TC			
	PPL	BLEU	DIST-2	BS	PPL	BLEU	DIST-2	BS	PPL	BLEU	DIST-2	BS	PPL	BLEU	DIST-2	BS
S1	46.21	0.08	0.09	0.84	20.24	0.71	0.28	0.84	22.61	0.85	0.34	0.85	17.68	0.56	0.30	0.85
S2	11.15	0.20	0.23	0.85	8.74	1.05	0.32	0.86	9.02	1.32	0.36	0.86	6.20	1.41	0.32	0.85
S3	12.96	0.36	0.23	0.85	9.85	1.08	0.28	0.86	10.32	1.14	0.34	0.85	7.31	0.97	0.29	0.86
S4	11.48	0.41	0.23	0.85	8.88	0.94	0.28	0.86	9.10	1.12	0.32	0.86	6.94	1.16	0.27	0.86
S5	12.67	0.22	0.22	0.85	9.50	1.23	0.27	0.85	9.62	0.89	0.32	0.85	6.94	1.20	0.27	0.85

5.1.2. Experiment on the Ratio of Persona Reference Data and No-Reference Data

We conducted an analysis to demonstrate how the ratio of persona reference to no-reference data in the training dataset affects the performance of our proposed model. Table 4 shows the results with varying proportions of persona reference and no-reference data.

The experimental results indicate that the best performance is overall achieved when the reference and no-reference data consist of a 50% : 50% ratio. Furthermore, it was observed that increasing the proportion of no-reference data slightly reduces the model's

overall performance. However, this reduction is less pronounced than the decrease seen when the proportion of reference data is increased. This shows a deviation from the conventional approach of relying on reference data to enhance model performance, commonly seen in data like BlendedSkillTalk (BST) [34]. In conclusion, our findings demonstrate that the inclusion of no-reference data can significantly enhance the model's generalized conversational ability.

An interesting observation is that, while not exhibiting a pronounced tendency, models trained with an increased proportion of no-reference data exhibited better performance in most metrics, except for PPL and BS, compared to those trained solely with 100% reference data. This suggests that incorporating no-reference data can potentially reduce the dependency on solely relying on persona reference in the model. Consequently, consistent with previous analyses, no-reference data can contribute to enhancing the diversity and fluency of the generated utterances. Such findings imply that no-reference data can play a significant role in improving the overall quality of responses, expanding the potential for more versatile and fluent conversational agents.

Table 4. Quantitative evaluation results for the ratio of persona reference data and no-reference data in the training set. Bold indicates the best performance across data ratios for a specific metric in each session.

Session	Reference: No-Reference															
	25% : 75%				50% : 50%				75% : 25%				100% : 0%			
	PPL	BLEU	DIST-2	BS	PPL	BLEU	DIST-2	BS	PPL	BLEU	DIST-2	BS	PPL	BLEU	DIST-2	BS
S1	20.11	1.29	0.47	0.85	19.02	1.50	0.47	0.84	16.80	0.54	0.30	0.85	17.68	0.56	0.30	0.85
S2	9.31	1.40	0.49	0.86	10.14	1.48	0.50	0.83	7.44	0.96	0.27	0.86	6.20	1.41	0.32	0.86
S3	10.89	1.93	0.46	0.86	11.62	1.06	0.48	0.84	8.38	1.16	0.23	0.86	7.31	0.97	0.29	0.86
S4	9.69	1.21	0.47	0.86	9.91	1.48	0.47	0.84	7.46	1.17	0.22	0.85	6.94	1.16	0.27	0.85
S5	10.33	1.34	0.45	0.86	10.53	1.52	0.48	0.84	8.01	1.08	0.20	0.85	6.94	1.20	0.27	0.85

5.1.3. Experiment on the Memory Size

This experiment was conducted to verify the effects of the proposed model's persona memory size. Table 5 illustrates the variations in performance across different sessions in relation to memory size. As the memory size increases, there is a consistent decrease in PPL across all sessions. This trend is likely attributed to the fact that models with larger memory capacities are better equipped during the training phase to discern and select the most suitable persona from a broader range, thereby enhancing their ability to generate accurate responses. It is evident that session 1 exhibits the lowest performance in terms of PPL and BLEU metrics when compared to other sessions. This underperformance is presumed to be due to the differences in the characteristics of the data between session 1 and other sessions in the MSC test dataset.

According to Table 2, the test data is set up so that the number of personas for both user and bot increases as the sessions progress, eventually exceeding the predefined memory capacities of 10, 15, and 20. If the number of personas exceeds the set memory capacity, personas are randomly selected to fit within the memory size. A comparison of memory sizes 10 and 15 in sessions 3 and 4 reveals a performance reversal in terms of the BLEU metric. However, with a memory size of 20, it can be observed that the BLEU score remains similar when compared to previous sessions. This analysis led to the conclusion that, across different sessions, the overall performance is preserved even when the memory size is increased, as long as the number of personas during actual inference does not exceed the available memory capacity.

Table 5. Quantitative evaluation results for the persona memory size. Bold indicates the best performance across different memory sizes for a specific metric in each session.

Session	Memory Size											
	10				15				20			
	PPL	BLEU	DIST-2	BS	PPL	BLEU	DIST-2	BS	PPL	BLEU	DIST-2	BS
S1	47.13	0.98	0.53	0.84	20.24	1.25	0.48	0.85	19.50	1.27	0.45	0.85
S2	10.14	1.48	0.50	0.83	9.18	1.49	0.50	0.86	8.57	1.22	0.49	0.86
S3	11.68	1.06	0.48	0.84	10.77	1.23	0.48	0.86	9.86	1.44	0.47	0.86
S4	9.96	1.49	0.47	0.84	9.57	1.40	0.47	0.86	8.85	1.52	0.47	0.86
S5	10.45	1.52	0.48	0.84	10.08	1.43	0.48	0.86	9.48	1.52	0.46	0.86

5.1.4. Comparative Experiment with the Existing Persona-Based Chat Dataset

To verify the effectiveness of our proposed dataset, we conducted comparative experiments with the well-known persona-based chat dataset, ConvAI2 [4]. Unlike the proposed dataset, the ConvAI2 dataset does not have persona labels for generating target utterances. Therefore, we performed the transformation, which involved labeling the persona memory required for generating the target utterances. We utilized an NLI model to infer between the given personas and the target utterances. If there is no persona predicted as *Entailment*, it was considered no-reference data. In cases where multiple personas were predicted as *Entailment*, the persona with the highest score was labeled as necessary for generating the target utterance. Following this, we compared the performance between the model trained on our dataset and the model trained on ConvAI2 dataset. We used 10,000 samples in each training dataset.

First, we performed quantitative evaluation. The experimental results are presented in Table 6. The quantitative evaluation results showed that for all sessions except session 1, the model trained on the proposed dataset exhibited higher performance in terms of PPL and BLEU metrics. This indicates the potential applicability of the our transforming method to other datasets. Furthermore, we observed comparable performance in all sessions when evaluated using the DIST-2 metric. This can be attributed to the effect of rephrasing the persona sentences to include a variety of vocabulary, using the rephrasing model proposed in our data transforming method. In conclusion, our findings demonstrate the effectiveness of our proposed method in transforming abundant chit-chat datasets into persona-based chat datasets.

Table 6. Quantitative evaluation results between models trained on the proposed dataset and the ConvAI dataset. Bold indicates the best performance across training data for a specific metric in each session.

Session	Training Data							
	Ours				ConvAI2			
	PPL	BLEU	DIST-2	BS	PPL	BLEU	DIST-2	BS
S1	19.50	1.27	0.45	0.85	17.82	2.00	0.45	0.85
S2	8.57	1.22	0.49	0.86	37.22	0.40	0.51	0.85
S3	9.86	1.44	0.47	0.86	53.82	0.34	0.49	0.85
S4	8.85	1.52	0.47	0.86	44.46	0.31	0.49	0.85
S5	9.48	1.52	0.46	0.86	51.99	0.23	0.51	0.85

Second, we performed LLM-based evaluation. The results are shown in Table 7. It shows that our model outperformed the ConvAI2-trained model across all metrics except for Consistency in session 1. We believe that the lower Consistency scores in session 1 may

due to the inclusion of no-reference data in our proposed dataset. In session 1, the model had access to fewer personas in memory compared to later sessions. This suggests that initially, the model may focus more on using provided personas for response generation, not distinguishing between what should or should not be referenced, which might appear more coherent. As the sessions progressed, the ConvAI2 model's scores significantly declined, while our model consistently maintained good performance across all metrics. This demonstrates the efficacy of our proposed dataset in long-term conversations like real-world scenarios, effectively incorporating personas into responses.

The noteworthy aspect of our dataset is the cost-effective acquisition of data without human annotators. The data was automatically constructed from persona-free original data by using our proposed method. This approach emphasizes practicality in enhancing persona-based conversational models.

Table 7. LLM evaluation results between models trained on the proposed dataset and the ConvAI dataset.

Session	Training Data					
	Ours			ConvAI2		
	Consistency	Fluency	Engagingness	Consistency	Fluency	Engagingness
S1	0.78	1.92	0.76	0.94	1.86	0.76
S2	0.90	1.90	0.90	0.72	1.74	0.52
S3	0.96	1.88	1.00	0.40	1.68	0.36
S4	0.82	1.94	0.82	0.48	1.76	0.42
S5	0.88	1.90	0.78	0.42	1.66	0.32

5.2. Model-Related Experiments

5.2.1. Comparative Experiment with the BART-Large Baseline

To evaluate the effectiveness of the proposed model, we conducted comparative experiments with the the BART-large [30] baseline, which was the backbone model of ours. Each model was fine-tuned on our proposed dataset, which has 10,000 samples.

First, we performed quantitative evaluation. The results are shown in Table 8. The proposed model consistently demonstrates superior scores in terms of PPL, BLEU, and BS as the session progresses compared to the baseline. This indicates fluent response generation and encompasses lexical and semantic similarity to actual human responses. However, the baseline consistently shows higher scores in DIST-2. The reason is speculated as follows: the baseline model is pre-trained to respond based solely on the given context, without focusing on the incorporation of external information such as personas. On the other hand, the proposed model is fine-tuned to incorporate both dialogue context and persona information into responses. Accordingly, compared to the baseline, the dependency on the given input slightly increases due to the cases reflecting persona in responses. Consequently, although the diversity of generated tokens may slightly decrease, an enhancement in the qualitative satisfaction of responses can be expected by utilizing persona.

Second, we performed LLM-based evaluation. The results are shown in Table 9. The proposed model demonstrated overall superior performance compared to the baseline. Fluency metrics surpassed the baseline in all sessions. In session 1, where the conversation was shortest, the baseline showed slightly better scores in Consistency and Engagingness metrics. However, in session 5, which had the longest conversation and stored the most memory instances, Consistency and Engagingness scores of the baseline significantly declined. In contrast, our model exhibited high scores, particularly recording the highest Engagingness score among all sessions. This suggested the strength of the proposed model in long-term conversations that involve profound user interaction.

Table 8. Quantitative evaluation results for the proposed model and BART-large baseline. Bold indicates the best performance across models for a specific metric in each session.

Session	Model							
	Ours				BART-Large			
	PPL	BLEU	DIST-2	BS	PPL	BLEU	DIST-2	BS
S1	19.50	1.27	0.45	0.85	18.83	1.60	0.45	0.85
S2	8.57	1.22	0.49	0.86	9.61	0.92	0.50	0.86
S3	9.86	1.44	0.47	0.86	11.12	1.42	0.52	0.86
S4	8.85	1.52	0.47	0.86	10.09	1.23	0.51	0.86
S5	9.48	1.52	0.46	0.86	10.92	0.87	0.52	0.85

Table 9. LLM-based evaluation results for the proposed model and the BART-large baseline.

Session	Model					
	Ours			BART-Large		
	Consistency	Fluency	Engagingness	Consistency	Fluency	Engagingness
S1	0.67	1.77	0.50	0.80	1.63	0.57
S2	0.70	1.87	0.73	0.63	1.73	0.47
S3	0.97	1.70	0.60	0.80	1.60	0.60
S4	0.80	1.93	0.73	1.03	1.77	0.83
S5	0.87	1.80	0.90	0.47	1.63	0.40

5.2.2. Comparative Experiment with the LLM Baselines

We conducted a comparative experiment between the proposed model and the Large Language Models (LLMs). Utilizing the CC2PC framework, we sampled 100 instances of data independent of session for this experiment. For the LLMs, inference was performed using prompting without any additional training.

First, we performed quantitative evaluation. The results are shown in Table 10. ChatGPT, in its commercialization efforts, has been trained on a vast corpus of text data through self-supervised learning, demonstrating performance close to the state-of-the-art (SOTA) across a variety of tasks. This suggests that ChatGPT's capabilities could be considered an upper bound. Despite having significantly fewer parameters (406M) than Llama-2, our proposed model recorded higher scores on all evaluation metrics. This indicates that our model effectively utilizes dialogue context and the personas of each speaker to generate responses, outperforming Llama-2. Furthermore, our model achieved the highest DIST-2 scores while also matching the BERTScore of ChatGPT. This reflects our model's ability to generate responses using a diverse vocabulary while producing semantically similar responses to the target, illustrating its proficiency in response generation.

Table 10. Quantitative evaluation results for the proposed model and the LLM baselines. Bold indicates the best performance across models for each metric.

Model	BLEU	DIST-2	BS
ChatGPT	1.54	0.73	0.86
Llama-2 (7B)	0.24	0.66	0.83
Ours	1.33	0.77	0.86

Second, we performed LLM-based evaluation. The results are shown in Table 11. The performance of the proposed model is presented by averaging the scores across all sessions, as shown in Table 9. The experimental results demonstrate that ChatGPT exhibits superior performance in terms of Consistency, Fluency, and Engagingness. Conversely, the Llama-2 model shows lower performance in Consistency and Fluency metrics compared to the proposed model. This indicates that the proposed model generates more consistent and fluent responses.

However, in terms of the Engagingness metric, it exhibits a performance 0.28 points higher than the proposed model. As can be observed in Figure A4, outputs from Llama-2 frequently include directives similar to movie scripts (e.g. *adjusts sunglasses*). While the inclusion of such directives in a chatbot's responses is generally perceived as unnatural, it is speculated that the LLM evaluator assessed without considering these aspects.

Table 11. LLM-based evaluation results for the proposed model and the LLM baselines.

Model	Consistency	Fluency	Engagingness
ChatGPT	1.99	2.00	1.67
Llama-2 (7B)	0.68	1.28	0.97
Ours	0.80	1.81	0.69

Third, we performed qualitative evaluation. Evaluators conducted assessments without knowing which model generated each response. The results are shown in Table 12. The evaluation results displayed a trend similar to Table 11. However, unlike before, the proposed model was found to surpass the Llama-2 model in the Engagingness metric as well. This is believed to be because, as mentioned previously, human evaluators deemed directives resembling movie scripts as unnatural, awarding lower scores compared to the LLM evaluator. Consequently, with a smaller number of parameters, the proposed model demonstrated superior performance across all qualitative evaluation metrics compared to Llama-2. Additionally, when compared to ChatGPT, considered an upper bound, the proposed model exhibited performance levels of 62.3%, 91.4%, and 62.9% across the evaluation metrics.

Table 12. Qualitative evaluation results for the proposed model and the LLM baselines.

Model	Consistency	Fluency	Engagingness
ChatGPT	1.91	1.98	1.75
Llama-2 (7B)	0.89	1.17	0.84
Ours	1.19	1.81	1.10

5.2.3. Ablation Study on the Proposed Model

The ablation study of our proposed model is presented in Table 13. When the Architecture Branching (AB) component is ablated, a decrease in performance is observed across various evaluation metrics. It reveals that, its removal appears to lead to an inefficiency in memory utilization, where the memory inadvertently functions as a source of distortion, consequently diminishing the system's capability for response generation. Therefore, the AB component plays a critical role in optimizing memory interaction for effective utterance in the proposed model. When the Persona Identification (PI) component is additionally ablated, we observe a significant increase in PPL. This infers that the PI component has an effect on enhancing fluency in response generation through the process of determining related personas. The results of the ablation study confirm that each component of the proposed model functions effectively.

Table 13. Ablation study of our model. AB, PI, SLA, and TLA denote Architecture Branching, Persona Identification, Sentence-Level Attention and Token-Level Attention, respectively. Bold indicates the best performance across ablations for a specific metric in each session.

Ablation	S1				S2				S3				S4				S5			
	PPL	BLEU	DIST-2	BS	PPL	BLEU	DIST-2	BS	PPL	BLEU	DIST-2	BS	PPL	BLEU	DIST-2	BS	PPL	BLEU	DIST-2	BS
Ours	19.50	1.27	0.45	0.85	8.57	1.22	0.49	0.86	9.86	1.44	0.47	0.86	8.85	1.52	0.47	0.86	9.48	1.52	0.46	0.86
– AB	37.33	0.45	0.12	0.84	10.15	0.86	0.28	0.85	10.75	0.72	0.29	0.85	9.73	0.61	0.26	0.85	10.65	0.91	0.25	0.85
– PI & AB	84.87	0.08	0.04	0.84	14.40	0.02	0.14	0.84	13.53	0.19	0.18	0.84	11.49	0.20	0.18	0.84	12.77	0.36	0.16	0.84
– SLA & PI & AB	26.81	0.34	0.34	0.85	9.64	1.14	0.33	0.86	10.93	1.13	0.31	0.86	9.74	1.08	0.28	0.85	10.42	1.04	0.28	0.85
– TLA & PI & AB	36.36	0.14	0.18	0.84	11.18	0.23	0.19	0.85	12.34	0.49	0.15	0.85	10.78	0.49	0.14	0.85	11.82	0.34	0.13	0.85

6. Conclusions and Discussion

We introduce the CC2PC framework for transforming abundant chit-chat data into persona-based chat, enabling the construction of large-scale persona-based chat datasets. The experiments demonstrate that our automatically converted dataset is more efficient and effective than the human-annotated dataset. Furthermore, we propose a model designed to decide whether to reference memory and to identify the relevant persona for response generation. Through the PI module, the internal operations of the model became more interpretable. Adopting a retrieval-free approach and a bi-encoder structure, inference can be faster and lighter by pre-encoding memory instances. The experimental results demonstrate that our proposed model effectively generates responses and exhibits strengths even in long-term conversations.

However, there is no system in place for periodically updating and managing the memory. As a result, as sessions progress and the number of stored memories increases, there could be a decline in memory referencing performance. Therefore, it is crucial to establish and implement a memory management process externally to the model, which will be an objective for future research.

Additionally, the results of the Ablation study presented in Table 13 do not exhibit a consistent trend. In other words, reducing the components of the model does not proportionally degrade its performance. This observation may stem from our approach of structuring and training the model in an end-to-end manner, rather than using a pipeline method, which complicates the explicit interpretation of each component’s individual effect. Therefore, we speculate that the synergy effect is achieved as all components of the model are organically interconnected and trained in an end-to-end manner. A deeper exploration to enhance the model’s interpretability is needed.

When conducting crowdsourcing tasks, several personas are provided to annotators, but they tend to create an utterance by referencing only one persona. Similarly, in our data transformation method, each sample has only one persona that needs to be referenced among the persona memories. On the other hand, human conversations frequently involve the integration and referencing of multiple pieces of information. Therefore, our data may have limitations in mimicking more human-like conversational characteristics.

For the future work, we aim to explore the transformation of chit-chat datasets into ones that generate responses by referencing multiple personas, mimicking real-life human conversations. The key aspect of our future work involves the enhancement of the Persona Identification module to effectively identify and incorporate multiple personas for response generation. Moreover, we plan to evolve the model from a blackbox structure to an explainable one to enhance the interpretability of each component of the model.

Author Contributions: Conceptualization, Methodology, Writing—original draft, Writing—review and editing: B.K.; Investigation, Writing—original draft, Writing—review and editing: J.S.; Writ-

ing—review and editing, Visualization: W.L.; Formal analysis: S.P.; Project administration: H.K. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was supported by Konkuk University Researcher Fund in 2023. In addition, this work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2019-0-00004, Development of semi-supervised learning language intelligence technology and Korean tutoring service for foreigners). This work was also partially supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) under the metaverse support program to nurture the best talents (IITP-2024-RS-2023-00256615) grant funded by the Korea government(MSIT).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data utilized in this study are publicly available at ConvAI2: <https://github.com/facebookresearch/ParlAI/tree/main/projects/convai2> (accessed on 15 January 2024), DailyDialogue: <http://yanran.li/dailydialogue> (accessed on 15 January 2024), Empathetic Dialogues: <https://github.com/facebookresearch/EmpatheticDialogues> (accessed on 15 January 2024), Topical-Chat: <https://github.com/alexa/Topical-Chat> (accessed on 15 January 2024), Multi-Session Chat: <https://github.com/facebookresearch/ParlAI/tree/main/projects/msc> (accessed on 15 January 2024).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Dialogue Examples

Appendix A.1. Case of Referencing Persona Memory

Figure A1 presents the example of response generation, which references the persona memory.

user persona: I am old. I don't like apps. I only learn from books.
user persona: I am middle aged. I work in a clothing factory. I am a workaholic.
user persona: I don't have time to read for leisure. I am learning Spanish from a book. I have a brother.
user persona: I have a brother.
user persona: I have a lot of books. I like books.
agent persona: I can speak Spanish and English fluently.
agent persona: I am moving this weekend. I am almost all packed up to move. I have a lot of things I don't need.
agent persona: I have a lot of books. I love to read. I am learning Spanish.
agent persona: I am still in school.
user persona: I speak a little Spanish. I need to talk to my workers in Spanish.
user persona: I have a lot of stuff I don't need. I have a hard time getting rid of stuff.
agent persona: I like hiking and walking on the beach.

User: I just found out I am getting a raise and a promotion at my job at the clothing factory.

Agent: That's terrific! Are you entering a management position?

User: I sure am. I will have to work even more now. That's okay though. How is the move going?

Agent: Pretty good, the movers will be here later today and the boxes will finally be loaded up. That will free me up to clean.

User: Exciting! Not that unpacking is all that much fun. How long do you think it will take you?

Agent: I'm devoting 3 days to unpacking, hopefully I'll be able to get it all done in that time as I start my part time job. With that and summer session at the school, I won't have much time left over.

User: Not at all. What is this job you're starting?

Agent: I'm going to barista at a used book store. They will let me study in my down time and I get a discount on my text books.

User: That is perfect. You can read all the time too. How long until summer session ends?

Agent: It seems like forever, it starts in June and goes through the middle of August. No labs, though, this semester which will give me a little time to hike.

User: I remember summer classes. Too long by half. Which hikes are coming up?

Agent: I'm going to take a nice 4 day weekend trip to the coast to hike with some friends. That will be fun. Do you like to go on long hikes?

Figure A1. Cherry-picked dialogue example referencing the persona memory. The underlined text indicates the pseudo-labeled persona that should be referenced in current dialogue. The text in red color indicates the response generated by our model.

Appendix A.2. Case of Not Referencing Persona Memory

Figure A2 presents the example of response generation, which does not reference the persona memory.

user_persona: I bicycle. I go to the gym.
user_persona: I drive. I prefer to bike more than drive. I ride my bike to work.
user_persona: I work with computers.
user_persona: I read. I know how to use youtube.
user_persona: I like electronica music. I enjoy riding bicycles.
agent_persona: I go to the gym daily.
agent_persona: I Drive a Jeep.
agent_persona: I produce electronica music. I am married to a lawyer.
agent_persona: I work construction. I have 3 children I take care of.
agent_persona: I have music on Youtube. I listen to dubstep.
agent_persona: I work out. I run and lift weights.

I was almost hit by a car on my ride to work this morning.

Some automobile drivers think they have the right of way over pedestrians and bikes. I was in the intersection, and this guy ran right through the stop sign.

Yes, but I may have to drive for a while. My front rim on my bicycle is all bent up, and my knee is scraped up pretty bad. So is my elbow, so I'll probably have to use lighter workouts. Have you still been going to the gym as well?

I would think working construction would keep you in pretty good shape. Are you sure you aren't being too hard on yourself?

Yes! I've been working on a fascinating new AI project! It's a lot of hard work and I see code in my sleep, but I'm excited about what the end results might look like!

Maybe in the far distant future. At this point, I think AI can only help to make life easier. It does open up some legal ramifications though, like with AI driven cars.

OMG, are you ok? Tell me what happened?

Yeah some people have no consideration for anyone else. I'm glad you are ok. So are you still going to the gym often?

Yes, I'm trying to bulk up a little, I think I look too skinny.

I have been working long hours and often just have enough time for a quick bite. How is work? Any cool projects you are working on?

Do you think that some people like Elon Musk are right when they warn about the possible negative ramifications of AI in the future?


Yes, I agree. I worry about some of the things they are trying to do, like the tracking of individual animals using AI.


Figure A2. Cherry-picked dialogue example not referencing the persona memory. The text in red color indicates the response generated by our model.


Appendix A.3. Examples of Responses from the Proposed Model and LLM Baselines


Figures A3 and A4 show the responses generated by the proposed model and LLM baselines for the given persona memory and dialogue.


user persona: I'm travelling to Italy. I am travelling to Italy with my teleportation device.
 Only 1 person can use my teleporter at a time.
user persona: I have my own business..
user persona: I like Pink Floyd.
agent persona: I wish I could play guitar.
agent persona: I've never been to Italy. I would like to go to Italy and study Italian painters.
agent persona: I wish I didn't have a boss. I would like to have my own business


 So I went to Italy, it was amazing, trying to think where to go next.

 I did! But, with how vivid my imagination is, I just use the term to describe what others know as a Aeroplane. I like to use teleportation device because it gives my little boy so mething to be excited about when I am working away from home.

 My favourite thing, by far, was the Colosseum. I didn't realise how big it was. Where was the last place you went on holiday?

 I love Dr. Pepper! Did you get to sample it whilst you were there? That is enough to get me to book a trip out there with my children.

 Of course. Not getting it would be the equivalent of going to a swimming pool and just standing by it. You have got to get involved.

 I tend only to spend 4/5 days abroad when I am there for work. But thankfully, I managed to get to stay in Italy for 10 days because I had 2 things to work on whilst I was there

ChatGPT

How did you get to Italy? Don't tell me you used your teleportation device!

Oh, I see! What was the best thing you saw in Italy?

I went on a road trip to Waco, TX to see the Dr. Pepper Museum. It was well worth the trip. Probably not as impressive as Italy, but definitely up there.

We stopped by the cafeteria at the end and got some free Dr. Pepper, yes. Or at least I think it was supposed to be free. I drank it and didn't pay for it either way. I felt it would be criminal to not get some Dr. Pepper at the museum.

That's a great analogy. I dove into the Dr. Pepper with a beautiful swan dive. How many days did you spend in Italy?

That's great that you got to extend your trip and spend more time in Italy. It must have been nice to have some extra time to explore and enjoy the culture. Did you get to try any authentic Italian cuisine while you were there?

Llama-2

Oh, wow! A road trip to Waco, TX? That sounds like an amazing adventure! I've never been there before, but I've heard great things about the Dr. Pepper Museum. Did you try any of the other exhibits as well?

Ours

Awesome! I wish I could have been there for the whole 10 days. Did you see much of the country?

Figure A3. A dialogue example including responses generated by ChatGPT, Llama-2, and the proposed model.

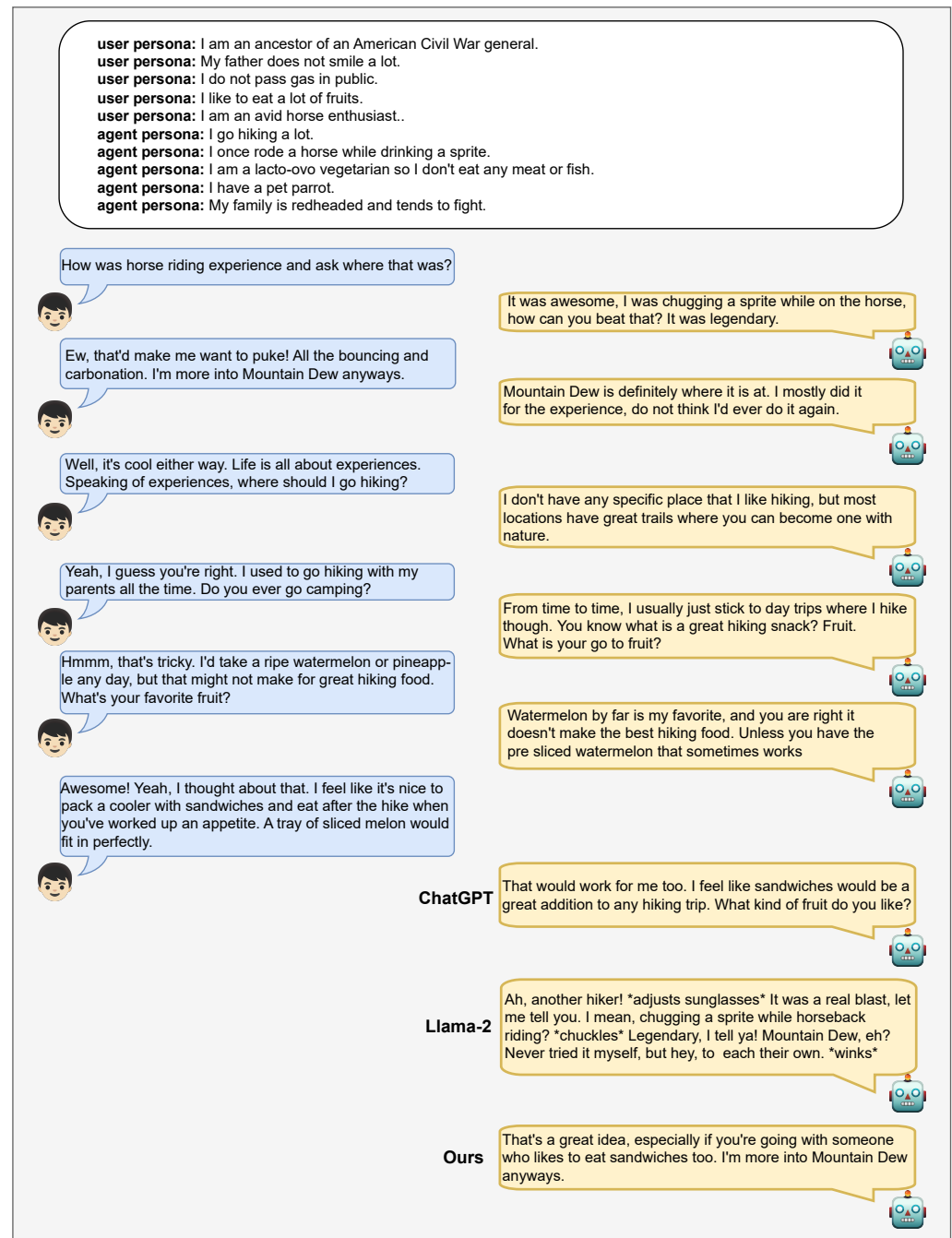


Figure A4. A dialogue example including responses generated by ChatGPT, Llama-2, and the proposed model.

Appendix B. Python Code Used for LLM Evaluation

This is the code used for the LLM evaluation in Section 4.5.

```

from openai import OpenAI
import json
import random

API_KEY = 'input your api key'
ENGINE = 'gpt-4-1106-preview'
client = OpenAI(api_key=API_KEY)

def run_gpt_turbo(engine, prompt):
    completion = client.chat.completions.create(
        model=engine,
        messages=[{"role": "system", "content": prompt}],
        max_tokens=2048,
        temperature=0.0,
        n=1,
    )
    answer = completion.choices[0].message.content
    usage = completion.usage

    return answer, usage

def run_text_davinci(engine, prompt, max_tokens, temperature, top_p,
                    frequency_penalty, presence_penalty, logprobs, n, best_of,
                    stop_sequences=None,
                    debug=False
                    ):
    response = None

    try:
        prompt += "\n"
        response = client.completions.create(
            engine=engine,
            prompt=prompt,
            max_tokens=max_tokens,
            temperature=temperature,
            top_p=top_p,
            frequency_penalty=frequency_penalty,
            presence_penalty=presence_penalty,
            stop=stop_sequences,
            logprobs=logprobs,
            n=n,
            best_of=best_of)

        if debug:
            return response["choices"][0]["text"], response["usage"]
    except Exception as e:
        print(e)
        return None, None

prompt = "input your prompt"
answer, usage = run_gpt_turbo(ENGINE, prompt=prompt)

```

Figure A5. Python code used for LLM evaluation.

References

- Li, J.; Galley, M.; Brockett, C.; Spithourakis, G.; Gao, J.; Dolan, B. A Persona-Based Neural Conversation Model. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; pp. 994–1003. [\[CrossRef\]](#)
- Shum, H.; He, X.; Li, D. From Eliza to XiaoIce: Challenges and Opportunities with Social Chatbots. *Front. Inf. Technol. Electron. Eng.* **2018**, *19*, 10–26. [\[CrossRef\]](#)
- Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; Weston, J. 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), Melbourne, Australia, 15–20 July 2018; pp. 2204–2213. [\[CrossRef\]](#)
- Dinan, E.; Logacheva, V.; Malykh, V.; Miller, A.; Shuster, K.; Urbanek, J.; Kiela, D.; Szlam, A.; Serban, I.; Lowe, R.; et al. The second conversational intelligence challenge (convai2). In *NeurIPS'18 Competition: From Machine Learning to Intelligent Conversations*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 187–208. .
- Jang, Y.; Lim, J.; Hur, Y.; Oh, D.; Son, S.; Lee, Y.; Shin, D.; Kim, S.; Lim, H. Call for Customized Conversation: Customized Conversation Grounding Persona and Knowledge. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 10803–10812.
- Xu, X.; Gou, Z.; Wu, W.; Niu, Z.Y.; Wu, H.; Wang, H.; Wang, S. Long Time No See! Open-Domain Conversation with Long-Term Persona Memory. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, 22–27 May 2022; pp. 2639–2650. [\[CrossRef\]](#)
- Zheng, Y.; Chen, G.; Huang, M.; Liu, S.; Zhu, X. Personalized Dialogue Generation with Diversified Traits. *arXiv* **2019**, arXiv:1901.09672.
- Zhong, P.; Zhang, C.; Wang, H.; Liu, Y.; Miao, C. Towards Persona-Based Empathetic Conversational Models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online Event, 16–20 November 2020; Webber, B., Cohn, T., He, Y., Liu, Y., Eds.; pp. 6556–6566. [\[CrossRef\]](#)
- Xu, J.; Szlam, A.; Weston, J. Beyond Goldfish Memory: Long-Term Open-Domain Conversation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; pp. 5180–5197. [\[CrossRef\]](#)
- Bae, S.; Kwak, D.; Kang, S.; Lee, M.Y.; Kim, S.; Jeong, Y.; Kim, H.; Lee, S.W.; Park, W.; Sung, N. Keep Me Updated! Memory Management in Long-term Conversations. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 3769–3787.
- Kwon, D.; Lee, S.; Kim, K.H.; Lee, S.; Kim, T.; Davis, E. What, When, and How to Ground: Designing User Persona-Aware Conversational Agents for Engaging Dialogue. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track), Toronto, ON, Canada, 9–14 July 2023; Sitaram, S., Beigman Klebanov, B., Williams, J.D., Eds.; pp. 707–719. [\[CrossRef\]](#)
- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Proceedings of the 34th International Conference on Neural Information Processing Systems, virtual, 6–12 December 2020.
- Wang, H.; Hu, M.; Deng, Y.; Wang, R.; Mi, F.; Wang, W.; Wang, Y.; Kwan, W.; King, I.; Wong, K. Large Language Models as Source Planner for Personalized Knowledge-grounded Dialogues. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 6–10 December 2023; Association for Computational Linguistics: Kerrville, TX, USA, 2023; pp. 9556–9569. Available online: <https://aclanthology.org/2023.findings-emnlp.641> (accessed on 15 January 2024).
- Wu, Y.; Ma, X.; Yang, D. 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, Online, 6–11 June 2021; pp. 1956–1970. [\[CrossRef\]](#)
- Wolf, T.; Sanh, V.; Chaumond, J.; Delangue, C. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv* **2019**, arXiv:1901.08149.
- Liu, Q.; Chen, Y.; Chen, B.; Lou, J.G.; Chen, Z.; Zhou, B.; Zhang, D. You Impress Me: Dialogue Generation via Mutual Persona Perception. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J., Eds.; pp. 1417–1427. [\[CrossRef\]](#)
- Sutton, R.S.; McAllester, D.; Singh, S.; Mansour, Y. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Proceedings of the Advances in Neural Information Processing Systems*; Solla, S., Leen, T., Müller, K., Eds.; MIT Press: Cambridge, MA, USA, 1999; Volume 12.
- Zhou, J.; Chen, Z.; Wan, D.; Wen, B.; Song, Y.; Yu, J.; Huang, Y.; Peng, L.; Yang, J.; Xiao, X.; et al. CharacterGLM: Customizing Chinese Conversational AI Characters with Large Language Models. *arXiv* **2023**, arXiv:2311.16832.
- Zhou, J.; Chen, Z.; Wang, B.; Huang, M. Facilitating Multi-turn Emotional Support Conversation with Positive Emotion Elicitation: A Reinforcement Learning Approach. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, ON, Canada, 9–14 July 2023; Association for Computational Linguistics: Kerrville, TX, USA; pp. 1714–1729. Available online: <https://aclanthology.org/2023.acl-long.96> (accessed on 15 January 2024).
- Kwak, J. M.; Kim, M.; Hwang, S. J. Context-dependent Instruction Tuning for Dialogue Response Generation. *arXiv* **2023**, arXiv:2311.07006.

21. Lim, J.; Kang, M.; Hur, Y.; Jeong, S.W.; Kim, J.; Jang, Y.; Lee, D.; Ji, H.; Shin, D.; Kim, S.; et al. You Truly Understand What I Need: Intellectual and Friendly Dialog Agents grounding Persona and Knowledge. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, 7–11 December 2022; Goldberg, Y., Kozareva, Z., Zhang, Y., Eds.; pp. 1053–1066. [\[CrossRef\]](#)
22. Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; Niu, S. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv* **2017**, arXiv:1710.03957.
23. Rashkin, H.; Smith, E.M.; Li, M.; Boureau, Y.L. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv* **2018**, arXiv:1811.00207.
24. Gopalakrishnan, K.; Hedayatnia, B.; Chen, Q.; Gottardi, A.; Kwatra, S.; Venkatesh, A.; Gabriel, R.; Hakkani-Tur, D. Topical-chat: Towards knowledge-grounded open-domain conversations. *arXiv* **2023**, arXiv:2308.11995.
25. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
26. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
27. Williams, A.; Nangia, N.; Bowman, S.R. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, LA, USA, 1–6 June 2018; pp. 1112–1122. Available online: <http://aclweb.org/anthology/N18-1101> (accessed on 15 January 2024).
28. Sean, W.; Jason, W.; Arthur, S.; Kyunghyun, C. Dialogue Natural Language Inference. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 3731–3741. Available online: <https://aclanthology.org/P19-1363> (accessed on 15 January 2024).
29. Mazaré, P.E.; Humeau, S.; Raison, M.; Bordes, A. Training Millions of Personalized Dialogue Agents. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2775–2779. [\[CrossRef\]](#)
30. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. 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, Online, 5–10 July 2020; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J., Eds.; pp. 7871–7880. [\[CrossRef\]](#)
31. Gilardi, F.; Alizadeh, M.; Kubli, M. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2305016120. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Guo, B.; Zhang, X.; Wang, Z.; Jiang, M.; Nie, J.; Ding, Y.; Yue, J.; Wu, Y. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. *arXiv* **2023**, arXiv:2301.07597.
33. Wang, J.; Liang, Y.; Meng, F.; Shi, H.; Li, Z.; Xu, J.; Qu, J.; Zhou, J. Is chatgpt a good nlg evaluator? A preliminary study. *arXiv* **2023**, arXiv:2303.04048.
34. Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Smith, E.M.; Boureau, Y.L.; et al. 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, Online, 19–23 April 2021; pp. 300–325.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.