# CS-433 Machine Learning - Project 2: Commonsense Persona-grounded Dialogue Challenge (CPDC)

Yihan Wang 352511
Zewei Zhang 351384
*EPFL, Switzerland*

***Abstract*—This study explores advancements in persona-based dialogue systems by integrating PEACOK, a Persona-grounded data set augmentation framework, with a Learn-to-Memorize-Entailment-and-Discourse-Relation (LMEDR) framework. The primary focus is on enhancing dialogue consistency and coherence by conditioning the model on enriched persona profiles and incorporating natural language inference for maintaining discourse structure. Utilizing augmented Persona-Chat and Dialogue-NLI datasets, we train the LMEDR model to produce more consistent and coherent dialogue responses. Our model shows comparable performance to GPT3.5 turbo with a simple prompt, suggesting its effectiveness. We also examine the impact of varying the number of extended personas and reaffirm the detrimental effect of including listener personas in training. A case study further validates the model's ability to generate coherent and persona-consistent responses.**

## I. Introduction

Traditional dialogue systems typically focus on achieving functional goals with precisely defined user intents, such as airline booking, restaurant reservations [1], and question answering [2]. However, generating meaningful and uninformative responses in a chit-chat setting remains a challenge due to the need for large datasets that capture the nuances of daily, piecemeal interactions.

Traditional chit-chat models face three major issues: (i) lack of a consistent personality, (ii) absence of explicit long-term memory, and (iii) a tendency to produce non-specific answers [3]. Zhang et al. [3] made significant strides in creating more engaging chit-chat models by introducing the PERSONA-CHAT dataset, which features configurable and consistent persona profiles. However, annotating persona-related datasets is costly, and given the complexity of real-world personas and the vast array of potential interactions, it is challenging for models to learn appropriate interactions from limited data alone. To address this, Gao et al. [4] introduced the Persona-grounded Commonsense Knowledge graph (PEACOK), which expands persona profiles using a knowledge graph framework and facilitates the discovery of interconnections between interlocutor personas.

Yet, maintaining consistency in dialogue involves more than just persona profiles. Discourse coherence is a key component of conversational effectiveness, encompassing the overall structure of the dialogue and the connections between utterances. Chen et al. [5] explored a BART-architecture-based method that incorporates latent entailment relations between premises and hypotheses, as well as discourse relations, introducing natural language inference (NLI) into persona-based dialogue.

In this study, we address the Commonsense Persona-Grounded Dialogue Challenge, hosted by AIcrowd and Sony. We explore two avenues for improvement on top of the LMEDR model [5]: performing various PEACOK [4] augmentations on the PERSONA-CHAT dataset [3] and conditioning the model on different persona settings by adjusting the LMEDR pipeline [5].

## II. Methods and Models

This section presents the methods and models we employ to enhance the consistency and coherence of persona-based dialogues. We make efforts primarily from two perspectives: the dataset perspective and the training model perspective.

### A. Persona-Consistent Dialogue Model

The primary objective of a persona-based dialogue model is to generate natural responses that align with a given persona. Previous studies have focused on incorporating persona embeddings into generative-based dialogue systems, such as the seq2seq model, as mentioned in Li et al.'s work [6]. The emergence of large-scale language models introduces a new approach

in terms of fine-tuning pre-trained models. Liu et al. proposed the $P^2$ BOT [7] to perceive and encode mutual personas, thereby enhancing dialogue coherence and consistency in the context of reinforcement learning.

Another approach to improve this model involves leveraging the Natural Language Inference (NLI) approach to construct latent features in a conversation. In one study, NLI is used to explicitly label the relationship between two elements (either persona or utterance) in a dialogue with three categories: entailment, neutral, and contradiction [8]. The specific definitions of these three labels are introduced in [8]. In essence, such categorization simplifies the modeling and training process for persona-based tasks.

We employ the Learn-to-Memorize-Entailment-and-Discourse-Relation (LMEDR) framework proposed by Chen et al. [5] for our challenge, considering its better performance in persona-based dialogue tasks compared to other frameworks. This model enhances consistency by acquiring the entailment relation memory with NLI for persona consistency and dialogue discourse memory for dialogue coherence. A pre-trained BART model is used to encode the text information and then convert it into latent embeddings. The details of training are available in the original paper by [5].

### *B. Persona-based Dialogue Dataset*

In line with the framework mentioned above, we utilize two datasets for training: Persona-Chat and Dialogue-NLI.

*1) Persona-Chat Dataset:* The Persona-Chat dataset comprises dialogues between paired speakers along with their profiles [3]. Each participant, whether a speaker or listener, is characterized by multiple profile sentences, as illustrated in Table I. Data in the Persona-Chat are categorized into three groups for our framework: the persona group, the query group, and the response group. Concatenating the content from these three groups forms the input for the BART encoder, which is used to train the component related to dialogue discourse memory.

Persona-Chat offers two types of datasets: the original and a revised version. The revised dataset consists of sentences that have been rewritten from the original, as shown in Table II. This revision aims to increase the difficulty of the task. The motivation behind creating this revised version is to address the issue of agents only repeating sentences they encountered in the training data, thereby promoting a more robust and adaptive dialogue model.

| Person 1 | Person 2 |
|---|---|
| I like to remodel homes. | I like canning and whittling. |
| I like to go hunting. | To stay in shape, i chase cheetahs at the zoo. |
| I like to shoot a bow. | In high school, i came in 6th in the 100 meter dash. |
| My favorite holiday is halloween. | I eat exclusively meat. |

Table I: Example of persona sentences in Persona-Chat data set

| Original version | Revised version |
|---|---|
| I like to remodel homes. | I love to redesign houses. |
| I like to go hunting. | Killing for sport is my hobby. |
| I like to shoot a bow. | I shot an arrow the other day! |
| My favorite holiday is halloween. | I like to get dressed up. |

Table II: Example of persona sentences in the original version and revised version

| Original | Augmented with PEACOK |
|---|---|
| I like to remodel homes. | I am a handyman, here is what i will do or achieve in the future, to renovate my house. |
| I like to go hunting. | I am a hunter, here is what i did in the past, went on a hunting trip. |

Table III: Example of augmented persona sentences with PEACOK

*2) Peacok:* As noted in Section I, PEACOK [4] enhances the agent's learning process by enriching the dataset with world knowledge from a knowledge graph. This knowledge graph in PEACOK is composed of three elements: persona, relation, and attribute, which correspond to the head, edge, and tail in a graph, respectively. Examples of augmented personas, as provided by PEACOK, are illustrated in Table III. We utilize PEACOK to augment the data in the PERSONA-CHAT dataset, subsequently training the model on this enriched dataset.

*3) Dialogue NLI Dataset:* The Dialogue NLI dataset, as introduced in Section II-A, is developed based on the principles of Natural Language Inference. We utilize this dataset to train the entailment relation memory component of the LMEDR model, specifically focusing on enhancing persona consistency. This training is essential for ensuring that the model's responses align accurately and consistently with the defined personas. Since the usage of this data set is different from the previous one, we call DNLI inference set (inference stage) and Persona-Chat training set (training stage) from now on to distinguish them.

| Model | Word F1 | BLEU |
|---|---|---|
| Ours | 17.270 | 0.867 |
| GPT3.5 (Simple Prompt) | 17.001 | **1.096** |
| BART (PeaCok) | **18.384** | 1.046 |

Table IV: Comparision of evaluation metrics with baseline models

## III. Results

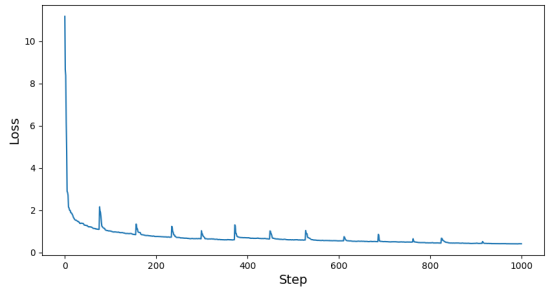
This section presents the training and evaluation results, and discusses the performance of our model. We conduct a comparative analysis across datasets augmented in various ways. This involves examining how much of data augmentation, such as the number of extensive personas with PEACOK, impact the effectiveness of the model. We assess the model's performance in terms of metrics like word F1, BLEU[9], accuracy, and perplexity (ppl) score to quantify the overall performance. The evaluation also includes comparisons with baseline models and previous approaches to highlight the effectiveness of our approach.

### A. Results of Our Best Model

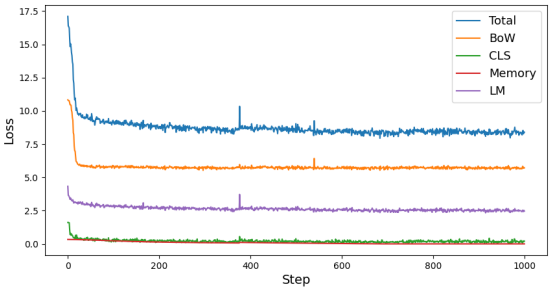
We obtained our best model by using input data that included augmented self-persona (speaker) information while excluding their-persona (listener) information from the original Persona-Chat dataset. The model was trained over 10 epochs with a scheduled learning rate. The batch sizes for the training and inference sets were 2 and 64, respectively. The convergent training curves for both the training and inference stages are shown in Figure 1. The spikes appearing in the plot are due to overlapping variables updated during both stages.

We also compared the evaluation results of our model with other baseline models from the CPD challenge, as illustrated in Table IV. The results show that our model surpasses the GPT-3.5-turbo model in terms of the Word F1 score when using a simple prompt. However, it falls short of the performance of the pretrained BART model [10], which was fine-tuned on the same augmented training set. Despite not outperforming the BART model, this comparison still demonstrates the effectiveness of our approach.

Due to the prolonged training duration (approximately 20 hours per epoch on SCITAS using two GPUs) and our limited time, it was impractical to train the model with datasets augmented in various ways, such as including only their-persona information




(a) Inference Loss




(b) Training Loss

Figure 1: Loss curve of our best model

or both personas in the training set. We opted to incorporate only the augmented self-persona into our long-training model. This decision was influenced by observations noted in [3] which suggest that dialogues in Persona-Chat, and most persona-based datasets, tend to focus more on the speakers themselves rather than the listeners. Consequently, we anticipated achieving a higher evaluation score under this specific setting.

### B. Impact of Number of Extensive Personas

As previously mentioned, the PEACOK framework extends personas with explicit relational information from the provided knowledge graph. The original paper [4] and the baseline model in CPDC set the maximum number of extensive personas at five, without exploring how model performance might vary with changes in this number. To investigate this, we augmented the datasets with varying maximum numbers of personas and compared the performance of models trained on these datasets. The adjustment of persona numbers was done by inducing both speaker (self) and listener (their) personas from the given reference list after each dialogue turn (one query-response pair), based on the dialogue history, and then retaining the most relevant

ones. Performance comparison metrics are based on accuracy and perplexity (ppl) scores on the validation set after completing the first epoch. The results are presented in Table V.

| Case | Max Number | Accuracy | PPL |
|---|---|---|---|
| 1 | 2 | **0.63** | **31.6** |
| 2 | 5 | 0.52 | 175.4 |
| 3 | 10 | 0.57 | 171.5 |

Table V: Comparision of evaluation metrics with baseline models

This time, we included both the original self-persona and added extra extensive self-personas and their-personas as input. Based on our findings, the model achieved its best performance — the highest accuracy score and the lowest perplexity — when limited to only two extensive personas. Indeed, this condition is similar with the scenario of including primarily self-personas in the data. This observation reaffirms our earlier assumption that incorporating their-persona into the training set does not enhance the model's performance when trained and evaluated with the Persona-Chat dataset. Besides, we also observed that increasing the maximum number of personas to 10 did not detrimentally affect performance.

### C. Case Study Analysis

To further assess our model's efficacy, given the limitations of automatic evaluation metrics for language tasks, we conducted a case study. We pre-established a dialogue history and then generated responses with our model based on the last query, using the same context as in [5]. The responses generated by our model, as well as by other models, are compiled in Table VI. Contrary to the unsuccessful attempts by the $P^2$BOT and BoB models, our combined model, along with the original LMEDR model, successfully generated satisfactory responses.

| | |
|---|---|
| Persona | I listen to rap music.<br>I produce music for artists.<br>I drive a 2015 honda civic.<br>My favourite food is pizza. |
| Context | Q: hi, how are you? do you have any brothers or sisters?<br>R: No i don't do you?<br>Q: yes , i'm 13 and i've an older brother.<br>R: that's nice what kind of music do you like<br>Q: i do not have much time as i play soccer. you?<br>R: i am a music producer for rap artists |
| Query | **cool i like rap .i hate maths though! do you have other hobbies** |
| GOLD | work takes up a lot of time |
| LIC | i love to eat pizza. |
| BoB | i like music and i like to listen to music (failed) |
| $P^2$BOT | i like to listen to rap music (failed) |
| LMEDR (without Peacok) | i like to drive my honda civic |
| GPT-3.5 Turbo simple prompt | i produce rap music, no other hobbies. |
| Ours | i like to eat pizza. |

Table VI: Case study of persona-based response generation using various models (generation results using other models refer to [5])

## IV. Conclusion

In this project, we integrated the PEACOK technique, a data augmentation framework designed for the Persona-Chat dataset, with a recently proposed, powerful persona-consistent dialogue generation method, LMEDR model. While our combined model did not surpass the performance of the BART model, it demonstrated comparable effectiveness to the GPT-3.5 turbo model when using simple prompts. Additionally, our investigation into the impact of the number of extended personas, as per the PEACOK framework, reaffirmed a performance decline when incorporating their-persona in the training process. The results from our case study further validated our model's capability to produce coherent and persona-consistent responses within the given context.

## V. Ethical Risks

Following the guidelines of the Digital Ethics Canvas, we conducted an early-stage assessment of potential ethical risks, as illustrated in Figure 2. Among the identified risks, the possibility of generating discriminatory responses based on real-world persona stereotypes is particularly concerning.

Training dialogue data, influenced by real-world biases, can inadvertently incorporate sexist or racist narratives, potentially leading the model to learn and generate offensive responses. Such biases can degrade user experiences and perpetuate existing discrimination. Unfortunately, this issue is not just theoretical; even advanced models like ChatGPT from OpenAI have been known to generate problematic responses, such as decisions based on nationality in contexts of torture.

Addressing this bias is most effective during the data collection phase. This can involve cross-examining narratives related to race, gender, health conditions, career, income level, etc., and revising or penalizing dis-

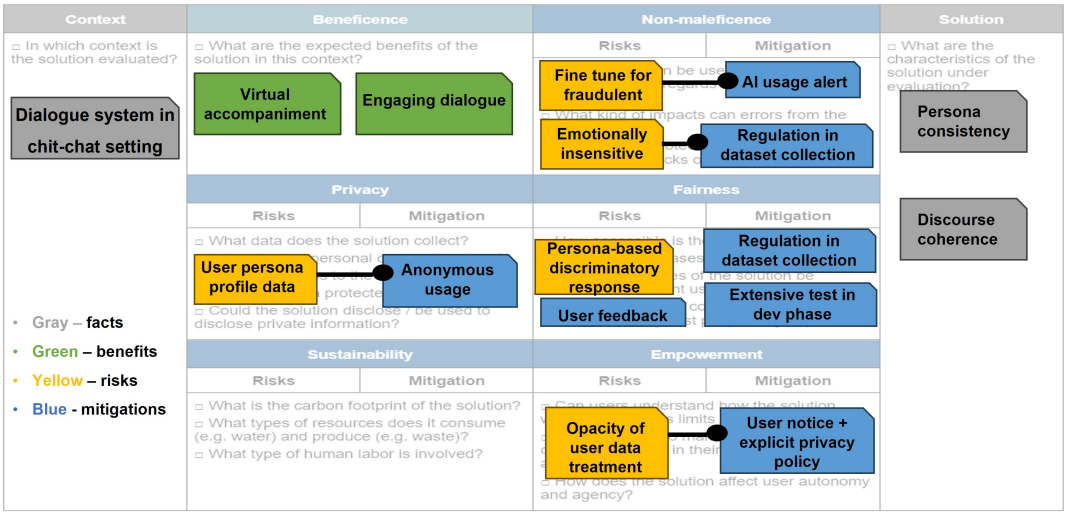



Figure 2: Ethical risk assessment

criminatory comments. However, the primary dataset used in our study, Persona-Chat [3], was collected with simple instructions like "get to know each other" and did not explicitly focus on ethical considerations. Since collecting and reviewing such data is costly and time-consuming, our project timeline did not allow for a comprehensive cross-examination, however, we would like to highlight this ethical risk to the Natural Language Processing research community.

## References


[1] A. Bordes and J. Weston, "Learning end-to-end goal-oriented dialog," *CoRR*, vol. abs/1605.07683, 2016. [Online]. Available: http://arxiv.org/abs/1605.07683

[2] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100, 000+ questions for machine comprehension of text," *CoRR*, vol. abs/1606.05250, 2016. [Online]. Available: http://arxiv.org/abs/1606.05250

[3] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing dialogue agents: I have a dog, do you have pets too?" *CoRR*, vol. abs/1801.07243, 2018. [Online]. Available: http://arxiv.org/abs/1801.07243

[4] S. Gao, B. Borges, S. Oh, D. Bayazit, S. Kanno, H. Wakaki, Y. Mitsufuji, and A. Bosselut, "Peacok: Persona commonsense knowledge for consistent and engaging narratives," 2023.

[5] R. Chen, J. Wang, L.-C. Yu, and X. Zhang, "Learning to memorize entailment and discourse relations for persona-consistent dialogues," 2023.

[6] J. Li, M. Galley, C. Brockett, G. P. Spithourakis, J. Gao, and B. Dolan, "A persona-based neural conversation model," 2016.

[7] Q. Liu, Y. Chen, B. Chen, J.-G. Lou, Z. Chen, B. Zhou, and D. Zhang, "You impress me: Dialogue generation via mutual persona perception," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 1417–1427. [Online]. Available: https://aclanthology.org/2020.acl-main.131

[8] S. Welleck, J. Weston, A. Szlam, and K. Cho, "Dialogue natural language inference," 2019.

[9] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. USA: Association for Computational Linguistics, 2002, p. 311–318. [Online]. Available: https://doi.org/10.3115/1073083.1073135

[10] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019.