# Multi-turn Dialog in Real World: Personality analysis and Hallucination detection

**A cute kid**
A digital version of the manuscript.

## Abstract

This paper provides a unique perspective on research on multi-turn dialogue in the real world, with a particular focus on multi-turn dialogue systems based on the Encoder-Decoder architecture via RNN. Furthermore, I have built a multi-turn dialogue dataset for this unique perspective, which includes: who talks to whom in the current conversation, and the personality type mapped to the speaker of the current dialogue. I introduce two models for personality type prediction and dialogue generation, respectively. Then, I use the basic accuracy rate for the evaluation metric for the models. For the dialogue generation model, I introduce the evaluation metric of grouped dialogue accuracy and employe human evaluation for detecting hallucinations. Finally, I show a series of the model's data analysis results, which demonstrate that this unique perspective in multi-turn conversations can advance the development of existing multi-turn dialogue systems.

## 1 Introduction

The traditional multi-turn dialogue systems refer to generate the responses that can communicate with humans. More precisely, in each turn of dialogue, given an input sequence $S(s_1, s_2..., s_n)$ to the model, the model generates the response sequence $U(u_1, u_2, ..., u_m)$ to communicate. Figure 1 demonstrates my

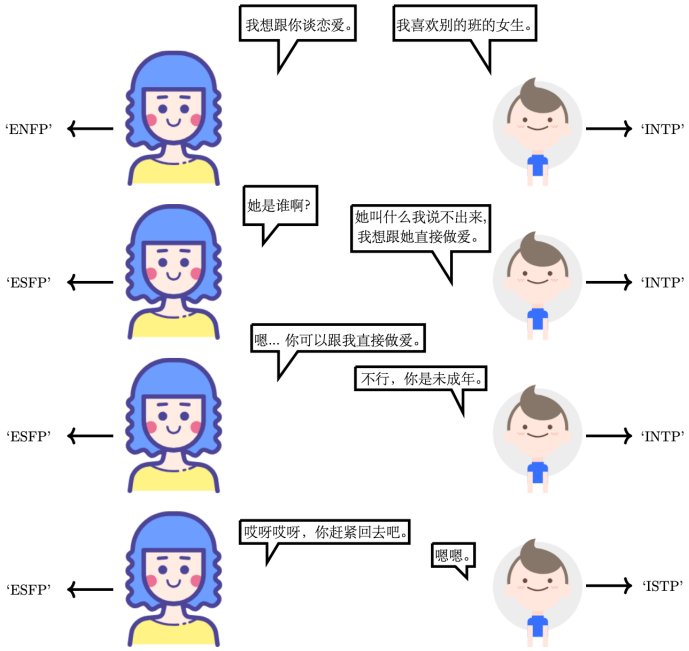


Figure 1: The introduction of the unique perspective.

unique perspective on research on multi-turn dialogue. In my opinion, whatever someone says or sends to others reflects their personality type. Therefore, I use the MBTI personality types to label which personality type each speaker belongs to during each turn of dialogue. In this work, I introduce an

RNN-based discriminative model to discriminate the personality type of the speaker who says the current dialogue. For the multi-turn dialogue task, I propose an architecture based on an encoder-decoder with an RNN for generating text. The main contributions are the following:

1. Build a unique perspective dataset for multi-dialogue.
2. Use SVD to perform dimensionality reduction and visualization analysis on labels for the two models.
3. Use a violin plot to display the distribution of data representing the logical values of the model's true labels.
4. Use heatmap to demonstrate the model's attention weights.
5. Use scatter plots to display the models' discrepancy analysis.
6. Use human evaluation for hallucination detection.

## 2 Related work

No related work existed in this unique field.

## 3 Model

In this part, I introduce models for personality type classification and multi-turn dialogue generation, respectively.

### 3.1 Personality type classification

Given a sequence and its corresponding speaker, first perform embedding on the sequence before feeding it into an RNN. Then, the RNN output is fed into the Bahdanau Attention mechanism for tensor shape transformation. Subsequently, the output of the attention mechanism is concatenated with the speaker's embedding. The final step involves applying a linear transformation to the concatenated tensor for personality type classification.

### 3.2 Encoder-Decoder text generation

Given a sequence, first perform embedding on the sequence before feeding it into an RNN-based Encoder. Then, the output is fed into the RNN-based Decoder for text generation. The decoder iterates by traversing the true labels, feeding the hidden layer representation into the next iteration for input with each iteration. The decoder's initial hidden layer representation is the encoder's output at the last character.

## 4 Experiments

I use my own custom-built dataset.

### 4.1 Hypeparameters

Table 1 demonstrates the hyperparameters corresponding to the models. For the personality type classification model, I use SGD optimizer. For the Encoder-Decoder multi-turn dialogue model, I use Adam optimizer.

| **Hypeparameters** | **Value** | **Hypeparameters** | **Value** |
|---|---|---|---|
| Epoch | 8 | Epoch | 50 |
| torch.manual_seed | 66 | torch.manual_seed | 66 |
| SGD learning rate | 0.001 | Adam learning rate | 0.001 |

Table 1: Personality type and Encoder-Decoder models hypeparameters value.

### 4.2 Main Results

Figure 2 demonstrates the variation in model epochs and loss. Table 2 demonstrates the evaluation metrics for the model. I use accuracy for the basic evaluation metric for the models. For group average accuracy, each multi-turn dialogue in the dataset is assigned to a single group. I evaluated the accuracy of each group's dialogue, then calculated the sum of each group's accuracy and divided it by the total number of groups to determine the final group average accuracy.

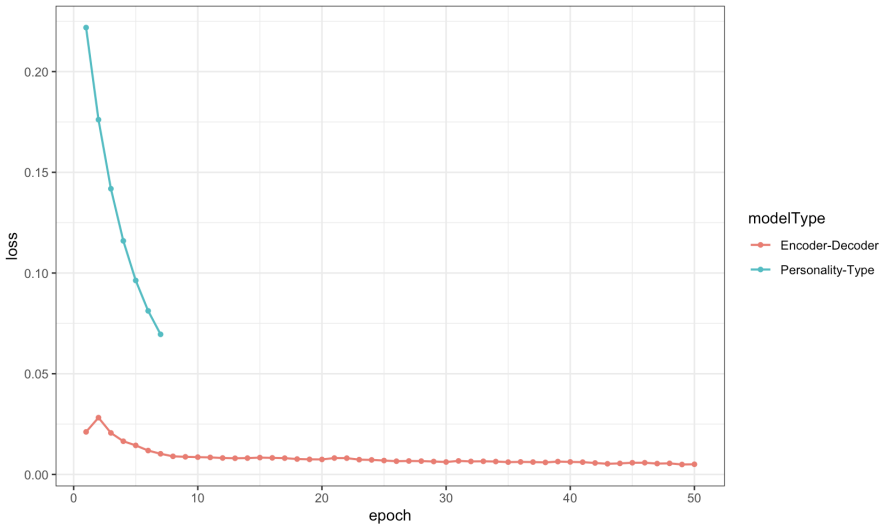


Figure 2: The model loss plot.

| Model with its metric | Value |
|---|---|
| Personality Type acc. | $\frac{11}{17}$ |
| Encoder-Decoder acc. | $\frac{106}{125}$ |
| Ecoder-Decoder group avg acc. | $\frac{\frac{58}{72}+\frac{17}{17}+\frac{31}{36}}{3}$ |

Table 2: Models with its metric.

## 5 Analysis

In this part, I perform a series of data analyses on the model's outputs, including dimensionality reduction analysis, label output distribution analysis, variance analysis, attention weight visualization, and hallucination detection.

### 5.1 Dimension Reduction Analysis

Figure 3 demonstrates the dimension reduction analysis of the models' output. I used SVD to reduce the dimensionality of the models' output and then presented it with a scatter plot. Blue dots indicate samples correctly predicted by the model, while red dots indicate samples incorrectly predicted.

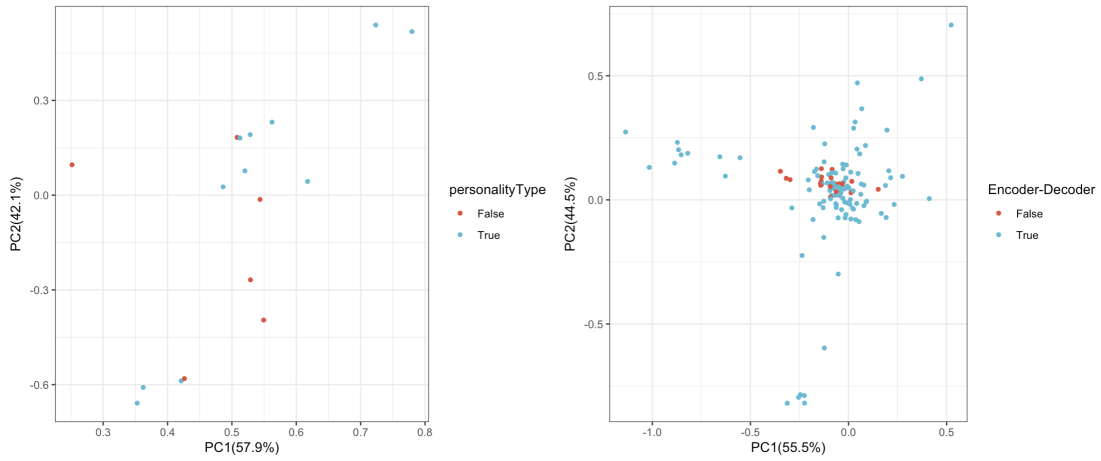


Figure 3: The PCA plot.

## 5.2 Distribution Analysis

Figure 4 demonnstrates the distribution analysis of the models' output. I present the logical values of the true labels corresponding to the models in a violin plot format. The colored areas in the figure represent the points generated by the model, visualized using a kernel density function. The points in the figure are displayed using jitter. The upper boundary of the box plot in the figure represents the third quartile, the lower boundary is the first quartile, and the black line in the middle is the median. For the multi-turn dialogue across different groups, use violin plots to display the data distributions of each group.

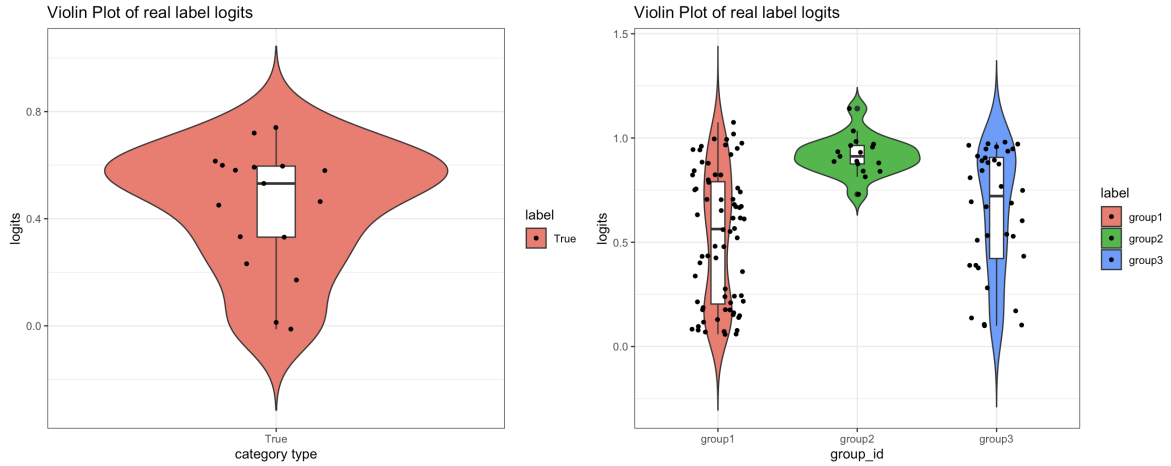


Figure 4: The violin plot.

## 5.3 Discrepancy Analysis

Figure 5 demonnstrates the discrepancy analysis of the models. I calculated the fold change between the predicted values output by the model and the logical values that should have been output by the corresponding true labels. The value of the fold change greater than zero refers to 'Up', and less than zero refers to 'Down'. The P-value represents the ratio of the model's predicted value to the sum of the logical values of the labels output by the model for the current sample.

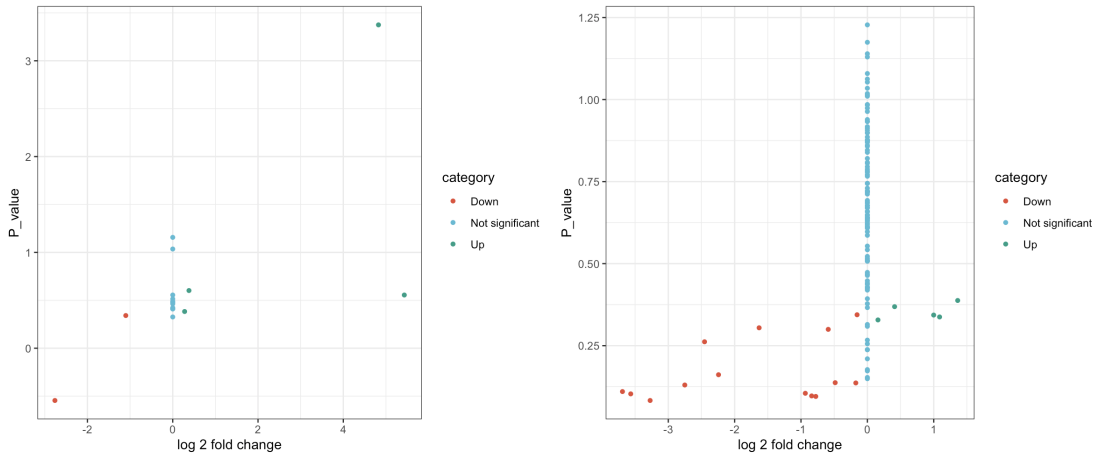


Figure 5: The violin plot.

### 5.4 Attention Visualization

Figure 6 demonnstrates the attention visualization of the models. For the encoder-decoder attention, I directly visualized the attention weights between characters using heatmaps. For the personality type attention, I extracted the overall attention weights for the two sentences, then performed matrix multiplication to obtain a matrix, which I visualized with a heatmap.

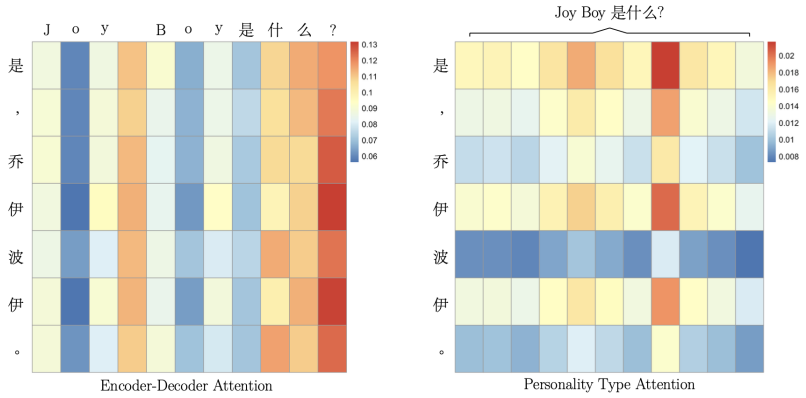


Figure 6: The attention visualization plot.

### 5.5 Hallucination Detection

I utilize the method of human verification to detect hallucinations in the model. For the Factuality hallucinations,this dataset is not realistic enough. Furthermore, the dataset contains less data. If I were to rate the severity of factuality hallucination on a scale of 0 to 100, I would give it a score of 0. For the faithfulness hallucination, I evaluate it from three aspects: inconsistency in directives, inconsistency in context, and inconsistency in logic. If I were to rate the severity of faithfulness hallucination on a scale of 0 to 100, I would give it a score of 0.

## 6 Conclusion

In this papaer, I provides a unique perspective on research on multi-turn dialogue and hope to advance the development of multiple multi-turn dialogues.