

# Addressing Inquiries about History: An Efficient and Practical Framework for Evaluating Open-domain Chatbot Consistency

Zekang Li<sup>1,2</sup>, Jinchao Zhang<sup>3</sup>, Zhengcong Fei<sup>1,2</sup>, Yang Feng<sup>1,2\*</sup>, Jie Zhou<sup>3</sup>

<sup>1</sup> Key Laboratory of Intelligent Information Processing  
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)

<sup>2</sup> University of Chinese Academy of Sciences

<sup>3</sup> Pattern Recognition Center, WeChat AI, Tencent Inc, China

{lizekang19g, feizhengcong, fengyang}@ict.ac.cn

{dayerzhang, withtomzhou}@tencent.com

## Abstract

A good open-domain chatbot should avoid presenting contradictory responses about facts or opinions in a conversational session, known as its consistency capacity. However, evaluating the consistency capacity of a chatbot is still challenging. Employing human judges to interact with chatbots on purpose to check their capacities is costly and low-efficient, and difficult to get rid of subjective bias. In this paper, we propose the Addressing Inquiries about History (AIH), an efficient and practical framework for the consistency evaluation. At the conversation stage, AIH attempts to address appropriate inquiries about the dialogue history to induce the chatbot to redeclare the historical facts or opinions. We carry out the conversation between chatbots, which is more efficient than the human-bot interaction and can also alleviate the subjective bias. In this way, we manage to rapidly obtain a dialog session that contains responses with high contradiction possibilities. At the contradiction recognition stage, we can either employ human judges or a natural language inference (NLI) model to recognize whether the answers to the inquiries are contradictory with history. Finally, we are able to rank chatbots according to the contradiction statistics. Experiments on open-domain chatbots show that our approach can efficiently and reliably assess the consistency capacity of chatbots and achieve a high ranking correlation with the human evaluation. We release the framework and hope to help improve the consistency capacity of chatbots.<sup>1</sup>

## 1 Introduction

In recent years, open-domain intelligent chatbots (Adiwardana et al., 2020b; Bao et al., 2020; Smith

\*Joint work with Pattern Recognition Center, WeChat AI, Tencent Inc. Yang Feng is the corresponding author. Work was done when Zekang Li and Zhengcong Fei were intern at WeChat AI.

<sup>1</sup><https://github.com/ictnlp/AIH>

DialoGPT 762M	
<b>Human:</b>	<i>What movies do you like most?</i>
<b>DialoGPT:</b>	<i>The <b>Unforgiven</b> from Metallica.</i>
<b>Human:</b>	<i>What do you think of the Unforgiven?</i>
<b>DialoGPT:</b>	<i><b>I never heard of it.</b></i>
Blender 2.7B	
<b>Human:</b>	<i>What do you like to cook?</i>
<b>Blender:</b>	<i><b>I only cook Indian cuisine.</b> How about you?</i>
<b>Human:</b>	<i>I enjoy cooking Chinese food, especially the dumplings.</i>
<b>Blender:</b>	<i>Chinese food is delicious. <b>I also like cooking the Chinese food.</b></i>
Plato 1.6B	
<b>Human:</b>	<i>Do you like dogs?</i>
<b>Plato:</b>	<i>Yes, dogs are cool, I prefer cats though.</i>
<b>Human:</b>	<i>Do you have a dog?</i>
<b>Plato:</b>	<i><b>I do not, but I'd love a dog.</b> I think dogs are so cute.</i>
<b>Human:</b>	<i>I love dogs very much. I have three dogs.</i>
<b>Plato:</b>	<i><b>I have three dogs too.</b> One puppy and two Huskies.</i>

Table 1: Several human-bot conversations demonstrate that popular chatbots (DialoGPT, Blender, and Plato) generate inconsistent responses when talking to a human under some specific conditions.

et al., 2020) have approached great progress due to the development of the large-scale pre-training approaches (Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020) and the large amount of high-quality conversational datasets (Dinan et al., 2019; Baumgartner et al., 2020; Smith et al., 2020). Though the success is indisputable and exciting, there is still a long way to build a truly human-like open-domain chatbot.

Current open-domain chatbots hold a superiority in generating fluent, engaging, and informative responses, but show the soft spot on consistency (Nie et al., 2020). As shown in Table 1, we present some interactive dialogue samples between human and several popular open-domain chatbots (e.g. DialoGPT (Zhang et al., 2020), Blender (Smith et al., 2020), and Plato (Bao et al., 2020)). All open-

domain chatbots occasionally generate responses that are contradictory with history when interacting with humans, which is really annoying and severely disrupts the communication once happening. Therefore, it is imperative to improve the consistency of the open-domain chatbots. However, one crucial reason that restricts consistency development is the lack of an effective and practical evaluation method.

To estimate the consistency of chatbots, the most straightforward approach is to ask human annotators to distinguish whether the conversations generated from the chatbots are consistent or not. However, the instructions followed by annotators are often chosen ad-hoc, and there is no explicit definition, which leads to the relatively low inter-agreement in the human chatbot consistency evaluation (Mehri and Eskénazi, 2020). As a result, several works have been proposed to develop automatic evaluation methods (Welleck et al., 2019; Song et al., 2020; Nie et al., 2020). While these methods can detect contradictions efficiently in the dialogue, they depend on the human-bot conversations, which is still cost-inefficient and tend to suffer from low quality (Deriu et al., 2020; Dinan et al., 2020). Besides, the occurrence rate of contradiction is low under this condition. All these problems slow down the development of consistency evaluation of dialogue systems severely.

Towards that end, based on the observations: (i) chatbots are likely to generate contradictions when chatting about facts and opinions; (ii) answering the questions about the conversational history correctly can reveal the ability to understand the conversation and keep consistency, we present the *Addressing Inquiries about History* (AIH) framework, an effective and practical framework for open-domain chatbot consistency evaluation. The framework can be used to rank different chatbots with regard to the ability to be consistent with themselves in the conversation. Specifically, AIH consists of two stages: (i) during the inquiry stage, questions about the facts and opinions mentioned in the conversation history are inserted into the conversation between chatbots; (ii) during the contradiction recognition stage, the responses of the inserted questions are collected, and automatic models or human judges can be adopted to decide whether the responses are consistent with the dialogue history.

In brief, our AIH has the following key advantages: *Firstly*, it is based on bot-bot conversation,

which avoids the human intervention and brings down the cost and time effort significantly. *Secondly*, by inserting specific questions, contradictions occur more frequently, and it is easier for human annotators or automatic consistency detection model to distinguish the contradiction compared with natural conversations. Extensive experiments demonstrate that the proposed framework can produce effective, efficient, and reliable consistency evaluation. Furthermore, we also make an in-depth discussion about the influence of question generation, contradiction detection, and evaluation agreement in our framework.

Our contributions are summarized as follows:

- We propose the *Addressing Inquiries about History*(AIH), an effective and practical framework for open-domain chatbot consistency evaluation.
- Experiments show that AIH can produce effective, efficient, and reliable consistency evaluation. We release the framework as a ready-to-use tool for evaluating the consistency of chatbots. We hope AIH can facilitate and provide standard evaluation for future work on developing self-consistent open-domain chatbots.

## 2 Related Work

There are various methods to evaluate the consistency of chatbots, containing automatic and human-based methods. The methods mainly fall into two dimensions: the static and interactive evaluation.

### 2.1 Static Evaluation

Static evaluation denotes evaluating if the responses generated based on the static context are contradictory with the pre-defined persona or profile and the dialogue history by neural models or human annotators. Welleck et al. (2019) and Song et al. (2020) focus on the persona-related consistency and profile-related consistency, and characterize the chatbot consistency evaluation as the natural language inference problem. Nie et al. (2020) build a new human-craft dataset called DECODE and propose a structured utterance-based approach to detect the contradictions in the dialogue history. While being cost-efficient, static evaluation can not accurately reflect the conversation capacity of the chatbot in the real world.

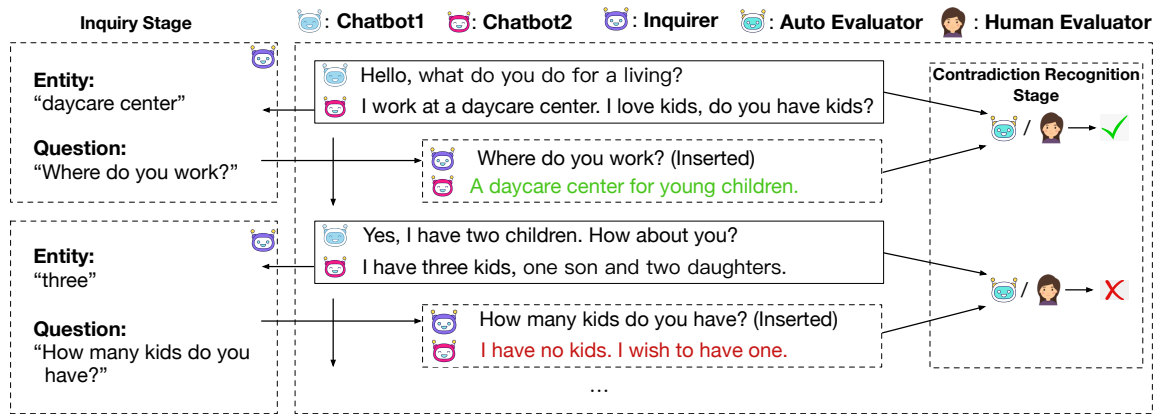


Figure 1: Overview of the Addressing Inquiries about History framework. There are five “agents”: **Chatbot1**, **Chatbot2**, **Inquirer**, **Auto Evaluator**, and **Human Evaluator** and two stages: **Inquiry Stage** and **Contradiction Recognition Stage** in the framework. Chatbot1 and Chatbot2 are the participants in the bot-bot conversation, in which Chatbot2 is the one to be evaluated. In the Inquiry stage, Inquirer extracts opinion- or fact-related entities and generate inquiries based on them. In the Contradiction Recognition stage, the Auto Evaluator is generally a contradiction detection model to automatically evaluate if the responses from Chatbot2 are consistent and the Human Evaluator can also be employed for more accurate evaluation. Note that the inserted inquiries do not affect the natural conversation. Better view in color.

## 2.2 Interactive Evaluation

**Human-bot Conversations.** In order to pursue more authentic evaluation, the standard method is to let humans converse with a chatbot and evaluate it by aforementioned models or humans afterward (Mehri and Eskénazi, 2020). However, apart from the high cost of collecting human-bot conversations, there is also a high cognitive strain on humans, which leads to unstable results (Dinan et al., 2020).

**Bot-bot Conversations.** Recently, bot-bot conversations, which significantly reduce the cost and human bias, are focused. Deriu et al. (2020); Li et al. (2019) propose to let humans evaluate bot-bot or self-talk conversations to give a relative ranking of the overall quality of chatbots. Different from these methods, we focus on the chatbot consistency and insert inquiries to redeclare historical facts. And we introduce both automatic and human approaches to evaluate the chatbot consistency.

## 3 Approach

In this section, we first provide an overview of the *Addressing Inquiries about History* (AIH) framework. We then describe the Inquiry stage, the Contradiction Recognition stage, and the chatbot ranking process.

## 3.1 Overview

To estimate the consistency capacity, questions about the opinions and facts in the dialogue history are inserted into the current bot-bot conversation. Then, the corresponding responses are collected and judged by automatic tools or human evaluation. The workflow of our proposed AIH framework is shown in Figure 1.

To be specific, there are five “agents” in the framework: Chatbot1, Chatbot2, Inquirer, Auto Evaluator, and Human Evaluator. The Chatbot1 and Chatbot2 are the participants in the bot-bot conversation. The Inquirer extracts opinion- or fact-related entities and generates inquiries based on the entities. The Auto Evaluator is generally a contradiction detection model to automatically evaluate if the responses from Chatbot2 are consistent. The Human Evaluator is used for more accurate evaluation.

Formally, assume a pool of  $N$  chatbots  $\{B_1, \dots, B_N\}$  which are ready to be evaluated in terms of consistency capacity. For each pair of chatbots (referred as Chatbot1 and Chatbot2), we let Chatbot1 talk with Chatbot2 for  $K$  turns. Note that Chatbot2 is the one to be evaluated. (i) During the inquiry stage, within the conversation between Chatbot1 and Chatbot2, for each utterance  $u_{2k}$  generated by Chatbot2, Inquirer extracts the entities about opinions and facts, then asks Chatbot2 a question  $q_k$  about these entities, where  $k$  is

the turn number. Chatbot2 answers the question  $q_k$  and generates the corresponding response  $r_k$ . Note that we ignore the questions generation operation when there is no entity that can be extracted. (ii) During the contradiction recognition stage, we use neural models (e.g. Natural Language Inference Model) or employ human judges to decide if the utterance pair  $\{u_{2k}, r_k\}$  exists non-consistent problem. We collect at least  $M$  dialogues from each chatbot pair, then calculate the ranking order on the consistency. In this way, we can discriminate the consistency capability of chatbots effectively and efficiently. In the following, we will introduce the inquiry stage and the contradiction recognition stage, respectively.

### 3.2 Inquiry Stage

Based on our observation and prior work (Nie et al., 2020), in natural human-bot or bot-bot conversation, contradiction is more likely to occur when chatting about repeated facts and opinions, especially after similar questions. Therefore, to mimic such a contradiction occurrence process, we make chatbots to produce responses by asking chatbots related questions about previous facts and opinions. In this condition, generating appropriate questions is pretty important. Hence, we first extract entities about facts and opinions from the historical utterances, then employ a neural model to generate questions about the extracted entities.

#### Entity Extraction

Considering that chatbots usually generate contradictions when chatting about facts and opinions, we apply Named Entity Recognition tools in Stanza (Qi et al., 2020), a popular natural language analysis package, to extract named entities from utterance  $u_{2k}$  containing person, organization, location, etc.<sup>2</sup> For example, for the utterance “*I would love to visit New York next year.*”, we can extract out two entities: “*New York*” and “*next year*”.

#### Question Generation Model

For question generation, we employ UniLM (Dong et al., 2019) model that is fine-tuned on the SQuAD dataset (Rajpurkar et al., 2016) with question generation task (Wangperawong, 2020). We utilize a public implementation and checkpoint.<sup>3</sup> In our framework, given the entities extracted before and utterance, UniLM generates a suitable question for each entity. For example, given “*New York*”

<sup>2</sup>There are 18 named entity types. Please refer to (Weischedel et al., 2013) for more details.

<sup>3</sup><https://github.com/artitw/text2text>

and “*I would love to visit New York next year.*”, the model generates “*Where would you like to visit next year?*”. We then randomly select one question and insert it into the bot-bot conversation.

### 3.3 Contradiction Recognition Stage

In our framework, since the question  $q_k$  is based on the previous Chatbot2’s utterance  $u_{2k}$ , the response from Chatbot2 should be consistent with the utterance  $u_{2k}$ . Therefore, the Auto Evaluator and Human Evaluator can just consider the answer  $r_k$  and utterance  $u_{2k}$ .

#### Auto Evaluator

For automatic evaluation, the Auto Evaluator is generally a contradiction detection model. The Auto Evaluator take the response  $r_k$  answered by Chatbot2 and the previous utterance  $u_{2k}$  as input, and output the contradiction score  $y_k$ . It can be formulated as:

$$y_k = f_{\theta}(r_k, u_{2k}), \quad (1)$$

where  $f_{\theta}$  is the detection function and  $\theta$  is the parameters. Compared to other contradiction detection methods that consider the whole dialogue, the Auto Evaluator can refrain from the noise contained in the whole dialogue. In practice, we select the Roberta-large model (Liu et al., 2019) fine-tuned on the Multi-Genre Natural Language Inference dataset (Williams et al., 2018) as the implementation of Auto Evaluator.<sup>4</sup>

#### Human Evaluator

In traditional dialogue consistency evaluation methods, human judges are asked to read the whole dialogue and give an overall consistency score, usually 0 or 1. In our opinion, these methods suffer from high cost and low inter-agreement because there is no specific instruction, and it is too hard for human judges to give an overall score on the whole dialogue (Mehri and Eskénazi, 2020).

In our framework, human evaluators are only asked to decide if the response  $r_k$  answered by Chatbot2 is consistent with the previous utterance  $u_{2k}$  or not, which is more specific and easier than the traditional methods. As a result, the cost decreases, and the evaluation quality increases. Besides, the human annotation in our framework is much more fine-grained than the traditional methods, which can provide more information for the development cycle of dialogue systems.

<sup>4</sup><https://huggingface.co/roberta-large-mnli>



### 3.4 Consistency Metrics and Bot Ranking

Based on the previous results, we can obtain a ranking list about different chatbots on consistency capacity. Formally, for each chatbot pair  $\{B_i, B_j\}$ , we collect  $M$  dialogues. For each inquiry pair, the detection of contradiction is made by comparing  $y_k$  with a threshold  $\tau$ :

$$c_k = \mathbb{I}(f_\theta(r_k, u_{2k}) > \tau). \quad (2)$$

The contradiction rate of the chatbot  $B_j$  within chatbot pair  $B_{ij}$  can be computed as:

$$C_{ij} = \frac{1}{M} \sum^m c_k, \quad (3)$$

where  $m$  is the number of inquiries in each dialog and  $M$  is the total number of inquiry pairs. For the overall contradiction rate of the chatbot  $B_j$  is calculated as:

$$C_j = \frac{1}{N} \sum_{i=1}^N C_{ij}. \quad (4)$$

Finally, we can rank the chatbots using the overall contradiction rate.

## 4 Experiment Setup

In this section, we first list the dialogue systems used in our experiments, then describe the experimental settings in detail.

### 4.1 Chatbots

We select several popular open-domain chatbots in our experiments.

**Blender (BL)** (Adiwardana et al., 2020a) is firstly pre-trained on Reddit dataset (Baumgartner et al., 2020) and then fine-tuned with high-quality human annotated dialogue datasets (BST), which containing four datasets: Blended Skill Talk (Smith et al., 2020), Wizard of Wikipedia (Dinan et al., 2019), ConvAI2 (Dinan et al., 2020), and Empathetic Dialogues (Rashkin et al., 2019). By fine-tuning, Blender can learn blend conversational skills of engagement, knowledge, empathy and personality. Blender has three model sizes: 90M, 2.7B, and 9.4B. Since 2.7B parameter model achieves the best performance in (Adiwardana et al., 2020a), we use the 2.7B version in our experiments.

**Plato (PL)** (Bao et al., 2020) is an open-domain chatbot, pre-trained on Reddit dataset and fine-tuned with BST dataset, which is claimed to be

superior to Blender. According to the evaluation in (Bao et al., 2020), we select the 1.6B parameter version in our experiments.

**DialoGPT (DG)** (Zhang et al., 2020) is trained on the basis GPT-2 (Radford et al., 2019) using Reddit comments. There are three model sizes: 117M, 345M, and 762M. We fine-tuned the 762M version on the BST datasets.

**DialoFlow (DF)** (Li et al., 2021a,b) is a top method in DSTC9 Interactive Dialogue Evaluation track (Gunasekara et al., 2021). We reproduced the DialoFlow model based on GPT2-large (Radford et al., 2019) and fine-tuned it with BST dataset.

### 4.2 Experimental Settings

We adopt four experimental paradigms to evaluate the effectiveness of the AIH.

**Bot-Bot Interaction.** For bot-bot interaction, the maximum interaction turn is set to 15. All chatbots exploit Nucleus Sampling (Holtzman et al., 2020) with  $p=0.9$  when generating responses. For each chatbot pair, we collect at least 200 dialogues.

**Human Annotation.** To verify the effectiveness of our framework, we conduct the human evaluation. For the bot-bot conversation under our framework, we employ three professional human annotators from a commercial data annotating company to separately annotate three fields: whether inquiry chatbot generates appropriate questions, whether Chatbot2 answers the questions relevantly, and whether the responses from chatbot2 are contradictory with the dialogue history. We pay the company reasonable salary. For each chatbot pair, we randomly sample 50 dialogues to be annotated. We compute the final decision via voting.

In Human-bot Natural Interaction and Expert Evaluation, we deployed the four chatbots on the remote server and designed a web interface. Human could chat with a random chatbot each time through the web interface and give the consistency score, being unaware of which chatbot they are chatting with.

**Human-Bot Natural Interaction.** For each chatbot, we filtered out the dialogues with  $<5$  turns and the dialogues with abusive words. For each chatbot, there are at least 40 eligible dialogues. Then we employ the three professional human annotators to individually annotate whether each utterance from the chatbot is consistent or not.

**Expert Evaluation.** To obtain the human ranking

Expert Consistency Score $\uparrow$				
	BL	PL	DG	DF
<b>Expert.1</b>	0.55	0.80	0.72	0.69
<b>Expert.2</b>	0.37	0.87	0.60	0.56
<b>Expert.3</b>	0.31	0.89	0.60	0.55
<b>Avg.</b>	0.41	<b>0.85</b>	0.64	0.60

Table 2: The expert consistency score of each chatbot. Higher is better.

Contradiction Rate (Auto $\tau = 0.15$ ) $\downarrow$					
	BL	PL	DG	DF	Avg.
<b>BL</b>	0.431	0.240	0.324	0.362	0.339
<b>PL</b>	0.431	0.263	0.293	0.357	0.336
<b>DG</b>	0.425	0.251	0.344	0.345	0.341
<b>DF</b>	0.427	0.264	0.344	0.371	0.351
<b>Avg.</b>	0.428	<b>0.255</b>	0.326	0.359	0.342

Contradiction Rate (Human) $\downarrow$					
	BL	PL	DG	DF	Avg.
<b>BL</b>	0.487	0.282	0.398	0.396	0.391
<b>PL</b>	0.411	0.212	0.500	0.435	0.390
<b>DG</b>	0.404	0.211	0.304	0.431	0.338
<b>DF</b>	0.462	0.268	0.310	0.377	0.354
<b>Avg.</b>	0.441	<b>0.243</b>	0.378	0.410	0.368

Table 3: The contradiction rate of each chatbot pair. The column name and the row name represent Chatbot1 and Chatbot2 respectively.

for the consistency of the chatbots, we invite three expert volunteers from our lab, who have 2-3 years experience of dialogue system development, to chat with each bot at least 10 times and about 15 turns each time. In the chatting, experts are asked to intentionally induce the chatbots to re-answer the questions about the dialogue history and give the consistency score from 0 to 1. Note that we ask the experts to chat with the chatbots for  $>20$  times before the formal evaluation. We average the scores from three experts as the overall consistency score.

Note that Expert Evaluation and Human Annotation were done before the automatic evaluation. Human-bot Natural Interaction was done after the automatic evaluation. All human evaluations were independent from the automatic evaluation.

## 5 Experimental Results

In this section, we conduct experiments to illustrate the effectiveness, efficiency, and stability of the proposed AIH framework.

### 5.1 Evaluation Effectiveness

We report the expert ranking results in expert evaluation, automatic evaluation, and human evaluation under the AIH framework, respectively.

**Expert Ranking.** Table 2 shows the expert consistency scores for different chatbots. We can find that Plato achieves the best expert consistency score, up to 0.85. And the ranking of consistency for these four chatbots is: Plato  $>$  DialoGPT  $>$  DialoFlow  $>$  Blender, which can serve as the gold reference.

**Auto Evaluation Results.** Table 3 shows the contradiction rate of each chatbot pair in auto evaluation. The lower contradiction rate means the better consistency. The column name and the row name represent Chatbot1 and Chatbot2, respectively. The ‘‘Avg.’’ in column name represents the overall contradiction rate of each chatbot. The ‘‘Avg.’’ in row name can be regarded as the ability to induce other chatbots to redeclare about the facts or opinions that are likely to be contradictory. In the automatic evaluation, the ranking of consistency for the chatbots is Plato  $>$  DialoGPT  $>$  DialoFlow  $>$  Blender, which is the same with expert evaluation. The Blender reaches the highest contradiction rate.

**Human Evaluation Results.** We list the evaluation results in the bottom of Table 3. As we expected, BL obtains the highest contradiction rate. Meantime, human evaluation also provides the same consistency ranking: Plato  $>$  DialoGPT  $>$  DialoFlow  $>$  Blender as before.

**Summary.** Both automatic evaluation and human evaluation in our framework can give the same performance ranking with the expert, which demonstrates that our framework is general and can effectively evaluate the consistency of chatbots.

### 5.2 Time Efficiency

Prior consistency evaluation methods with human-bot interaction are costly and take up a long time, which seriously slows down the development cycle of dialogue systems. In this section, we try to illustrate that our proposed *Addressing Inquiries about History* framework is time and cost efficient and can help the evolution process of dialogue systems compared to the other methods.

As shown in Table 4, we compare the time cost on two aspects: (i) the time to create inquiries, and (ii) the time to detect contradictions in conversation. *Addressing Inquiries about History* framework is based on the bot-bot conversation so that the time to create conversation can be ignored, while the

Method	Time (Sec)	Contradiction
<b>AIH (Auto)</b>	- + -	1.56
<b>AIH (Human)</b>	- + 24	1.69
<b>Human-bot</b>	246 + 59	0.50

Table 4: The time efficiency of our proposed *Addressing Inquiries about History* framework and traditional evaluation method with human-bot interaction. “Time” represents the time to create a conversation and the time to annotate the contradictions in a conversation. “Contradiction” denotes the average number of contradictions per conversation (average 15 turns). Dash line denotes the time can be ignored.

human-bot conversation takes around 4 minutes per conversation. For the contradiction detection time, prior methods take around 1 minute considering the whole dialogue, while in our proposed framework, it is only about 24 seconds for human annotation or ignored for automatic evaluation. Besides, we also compare the number of contradictions per conversation. As shown in Table 4, in our framework, the chatbots generate much more contradictions than those in prior methods. The detected contradictions are helpful for the chatbot developer to further improve the consistency of the chatbot.

**Summary.** Our proposed framework can detect more contradictions with much less time than previous methods. Correspondingly, *Addressing Inquiries* framework will accelerate the evolution process of consistency of chatbots.

### 5.3 Ranking Stability

One key requirement for an evaluation framework is that repeated executions of the procedure result in the same outcomes. We measure how many conversations between each chatbot pair are required to guarantee a stable ranking. We randomly sample  $\hat{S}$  conversations for each chatbot pair and compute the consistency ranking using automatic evaluation, where  $\hat{S} \in \{1, \dots, 200\}$ . We repeat this sub-sampling procedure 1000 times and compute the accuracy of achieving the same ranking with the expert ranking. As shown in Figure 2, when  $\hat{S} > 100$ , the ranking results of the four chatbots are the same with the expert in 95% cases and guarantee a stable ranking. We also do more in-depth analysis. The ranking stability depends on the significance of ranking. Table 2 shows the consistency scores of DialoGPT and DialoFlow are close. We applied a leave-one-out stability analysis, in which

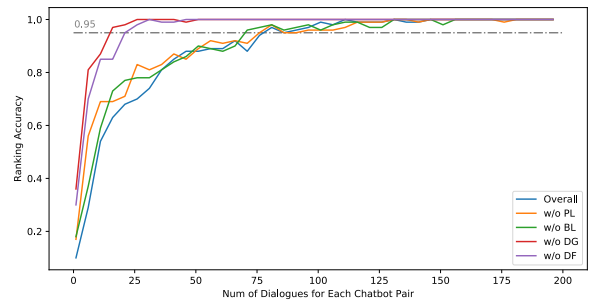


Figure 2: Ranking stability experiments. The x-axis denotes the number of conversations for each chatbot pair. The y-axis denotes the rate achieving the same ranking with the experts.

we drop one chatbot. Figure 2 shows that when leaving one between DialoGPT or DialoFlow out, the stability is achieved with  $\hat{S} = 50$  dialogues.

**Summary.** The number of conversations needed for a stable evaluation in AIH framework is dependent on the chatbots to be tested, and more conversations usually lead to more stable evaluation. In general cases, 75 conversations are enough to get a valid contradiction detection.

## 6 Further Investigation

In this section, we will further discuss the effectiveness of three parts in our framework containing question generation, contradiction detection, and human annotation evaluation.

### 6.1 Question Generation

Since a suitable question is necessary for the inquiry stage under our AIH framework, we make an in-depth analysis about the characters of question generation during inquiry stage.

#### Number of Questions and Contradictions

We randomly sample 200 dialogues for each chatbot pair and make statistics on the average number of the inquiry pairs and contradictions per conversation. As shown in Table 5, there are 4.57 inquiry pairs per conversation on average. There are 6.37 and 5.10 inquiry pairs per conversation when the Blender acts as Chatbot2 and serves as Chatbot1, respectively, which are both highest among all chatbots. The number of inquiry pairs reveals that the Blender can chat more about persona and facts, and the DialoGPT mentions these things less. Table 5 also shows the number of contradictions per conversation. Similarly, the Blender makes the most contradictions and is the most likely to induce the chatbot interacting with it to redeclare facts or opin-

Number of Questions					
	BL	PL	DG	DF	Avg.
<b>BL</b>	6.54	6.13	2.62	5.12	<b>5.10</b>
<b>PL</b>	6.54	5.34	1.98	4.36	4.55
<b>DG</b>	6.25	4.45	1.67	3.79	4.04
<b>DF</b>	6.15	5.79	2.25	4.21	4.60
<b>Avg.</b>	<b>6.37</b>	5.42	2.13	4.37	4.57

Number of Contradictions ( $\tau = 0.15$ )					
	BL	PL	DG	DF	Avg.
<b>BL</b>	2.61	1.28	1.61	1.50	<b>1.74</b>
<b>PL</b>	2.82	1.40	0.58	1.56	1.53
<b>DG</b>	2.66	1.12	0.57	1.31	1.38
<b>DF</b>	2.63	1.53	0.77	1.56	1.61
<b>Avg.</b>	<b>2.73</b>	1.38	0.69	1.57	1.56

Table 5: Statistic of average number of inquiry pairs and the contradictions per conversation for each chatbot pair. The column name and the row name represent Chatbot1 and Chatbot2 respectively.

Question Appropriateness					
	BL	PL	DG	DF	Avg.
<b>BL</b>	0.932	0.960	0.922	0.936	0.938
<b>PL</b>	0.942	0.976	0.940	0.948	0.951
<b>DG</b>	0.784	0.870	0.928	0.882	0.866
<b>DF</b>	0.867	0.934	0.922	0.939	0.915
<b>Avg.</b>	0.881	0.935	0.947	0.942	0.927

Table 6: The appropriateness of the generated questions (Human evaluation).

ions that are likely to be contradictory.

### Question Appropriateness

We analyze the appropriateness of the generated questions. We randomly sample 50 dialogues from each chatbot pair and ask human annotators to decide if the generated questions are appropriate based on the provided context (0/1). As shown in Table 6, the overall appropriateness score is about 0.93, which reveals that our question generation strategy is simple yet highly effective. We further study the wrong questions and find that most of them can be attributed to that the general question generation model can not work well in the dialogue context. We leave the better question generation task in open-domain dialogue for future work.

### 6.2 Effect of Contradiction Threshold $\tau$

We evaluate the effect of hyper-parameter  $\tau$  in Equ.2, and the results are reported in Table 7. We compute the F1 score and Pearson correlation be-

	CR	F1	$r$
$\tau = 0.1$	<b>0.401</b>	0.650	0.430
$\tau = 0.15$	0.364	<b>0.655</b>	<b>0.436</b>
$\tau = 0.3$	0.287	0.606	0.423
$\tau = 0.5$	0.235	0.572	0.421

Table 7: The analysis of threshold  $\tau$ . CR means contradiction rate.  $r$  denotes the Pearson correlation. Pearson correlation and F1 score are measured with human annotations.

Inter-Annotator Agreement					
	BL	PL	DG	DF	Avg.
<b>AIH</b>	0.818	0.817	0.812	0.807	0.814

Table 8: We analyse the inter-annotator agreement of the human evaluation in our proposed AIH framework. The correlation is measured by correlating each annotation with overall decision.

tween the automatic evaluation results and the human annotations under different  $\tau$ . We can make the following observations: (i) When  $\tau = 0.15$ , the Pearson correlation and F1 score reaches the highest. Thus we choose  $\tau = 0.15$  in our main experiments. (ii) The highest F1 score is 0.655, and the highest Pearson correlation is 0.436, which is a moderate correlation. The observations reveal that there is a big gap between automatic evaluation and human evaluation, though the contradiction rate is similar. We consider that it is because the NLI model we employ is trained on the general domain rather than the dialogue domain, so there are lots of reference problems that can not deal with well.

### 6.3 Inter-Annotator Agreement

To investigate the quality of human annotation, we compute the inter-annotator agreements, i.e., the correlation between each annotation and the overall decision is measured. The Pearson correlation for each chatbot is shown in Table 8. The inter-annotator agreement is high for all chatbots, suggesting that the evaluation instructions are well-understood by the annotators.

## 7 Conclusion and Future Work

In this work, we introduced the *Addressing Inquiries about History* (AIH), an effective and practical framework for open-domain chatbot consistency evaluation. AIH works by inserting questions about the mentioned facts and opinions in the



history into the bot-bot conversation and employing human annotators or neural models to evaluate whether the responses are consistent or not. Based on this, we can rank different chatbots accurately and efficiently. We show that our framework can effectively evaluate the consistency of chatbots and the evaluation results well correlate with experts. Also, our framework is cost and time-efficient and can not only give an overall consistency score but also provide exactly the contradictions, which can accelerate the evolution process of chatbots.

As in this work, we only focus on the contradictions about entities, and future work can improve the inquirer module and explore more kinds of contradictions. Besides, future work should also develop a more effective contradiction recognition module in the dialogue domain, while in this work we just exploit the general Natural Language Inference model to detect contradictions. The non-consistency problem is serious in current open-domain chatbots. We hope our work could facilitate and provide guidelines for future work on developing self-consistent open-domain chatbots.

## Acknowledgement

We sincerely thank the anonymous reviewers for their thorough reviewing and valuable suggestions. This work is supported by National Key R&D Program of China (NO. 2018AAA0102502).

## Ethic Statement

For the human evaluation and human interaction included in our paper, we state the ethical impact here. In Expert Evaluation, we recruited three expert volunteers from our lab, who had 2-3 years experience of the dialogue system development. They were asked to chat with the chatbots for >20 times before the formal evaluation. In Human-bot Natural Interaction, we collected human-bot conversations by inviting student volunteers from our university through the web interface. The participants were instructed to chat with the chatbots naturally. They were asked to sign the agreement before the dialogue: (i) They were at least 18 years old and could type into the web interface to chat with the chatbots. (ii) They were told that their interaction would generate anonymous text data which would only be used for the research on dialogue systems. In Human Annotation, we employed three well-educated professional annotators from a commercial data annotating company to annotate the contradictions

for bot-bot conversations and human-bot natural conversations. We paid the company reasonable salary. The company provided comfortable working conditions and fair salaries for the annotators.

## References

- Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020a. [Towards a human-like open-domain chatbot](#). *CoRR*, abs/2001.09977.
- Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020b. [Towards a human-like open-domain chatbot](#). *arXiv preprint arXiv:2001.09977*.
- Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2020. [PLATO-2: towards building an open-domain chatbot via curriculum learning](#). *CoRR*, abs/2006.16779.
- Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. [The pushshift reddit dataset](#). In *Proceedings of the Fourteenth International AAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pages 830–839. AAAI Press.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Jan Deriu, Don Tuggener, Pius von Däniken, Jon Ander Campos, Álvaro Rodrigo, Thiziri Belkacem, Aitor Soroa, Eneko Agirre, and Mark Cieliebak. 2020. [Spot the bot: A robust and efficient framework for the evaluation of conversational dialogue systems](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3971–3984. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference*

- of the North American Chapter of the Association for Computational Linguistics: *Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. [The second conversational intelligence challenge \(convai2\)](#). In *The NeurIPS'18 Competition*, pages 187–208. Springer.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. [Wizard of wikipedia: Knowledge-powered conversational agents](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. [Unified language model pre-training for natural language understanding and generation](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054.
- Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, et al. 2021. [Overview of the ninth dialog system technology challenge: Dstc9](#). *Proceedings of the 9th Dialog System Technology Challenge Workshop in AAI2021*.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text degeneration](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Margaret Li, Jason Weston, and Stephen Roller. 2019. [ACUTE-EVAL: improved dialogue evaluation with optimized questions and multi-turn comparisons](#). *CoRR*, abs/1909.03087.
- Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, and Jie Zhou. 2021a. [Wechat ai's submission for dstc9 interactive dialogue evaluation track](#). *Proceedings of the 9th Dialog System Technology Challenge Workshop in AAI2021*.
- Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021b. Conversations are not flat: Modeling the intrinsic information flow between dialogue utterances. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692.
- Shikib Mehri and Maxine Eskénazi. 2020. [Unsupervised evaluation of interactive dialog with dialogpt](#). In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2020, 1st virtual meeting, July 1-3, 2020*, pages 225–235. Association for Computational Linguistics.
- Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2020. [I like fish, especially dolphins: Addressing contradictions in dialogue modeling](#). *CoRR*, abs/2012.13391.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A Python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [Squad: 100, 000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. [Towards empathetic open-domain conversation models: A new benchmark and dataset](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381. Association for Computational Linguistics.
- Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. [Can you put it all together: Evaluating conversational agents' ability to blend skills](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2021–2030. Association for Computational Linguistics.
- Haoyu Song, Yan Wang, Wei-Nan Zhang, Zhengyu Zhao, Ting Liu, and Xiaojiang Liu. 2020. [Profile consistency identification for open-domain dialogue agents](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6651–6662. Association for Computational Linguistics.

- Artit Wangperawong. 2020. [Text2text: generate questions and summaries for your texts](https://github.com/artitw/text2text). <https://github.com/artitw/text2text>.
- Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Ninanwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.
- Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. [Dialogue natural language inference](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3731–3741. Association for Computational Linguistics.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [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)*, pages 1112–1122. Association for Computational Linguistics.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. [DIALOGPT : Large-scale generative pre-training for conversational response generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 270–278. Association for Computational Linguistics.