What would Harry say? Building Dialogue Agents for Characters in a Story

Nuo Chen[†,*], **Yan Wang**[‡§], **Haiyun Jiang**[‡]
Deng Cai[‡], **Ziyang Chen**[‡], **Longyue Wang**[‡] and **Jia Li**[†§]
[‡]Tencent AI Lab
[†]Hong Kong University of Science and Technology (Guangzhou),
Hong Kong University of Science and Technology
[†]chennuo26@gmail.com,[§] {yanwang.branden@gmail.com, jialee@ust.hk}

Abstract

We have a Christmas gift for Harry Potter fans all over the world. In this paper, we present Harry Potter Dialogue (HPD), a dataset that helps train Harry Potter-like dialogue agents. Such a task is typically viewed as a variant of personalized dialogue agents, but they differ significantly in three respects: 1) Harry lived in a virtual world of wizards, thus, real-world commonsense may not apply to Harry's conversations; 2) Harry's behavior is strongly linked to background information in conversations: the scene, its attributes and its relationship to other speakers; and 3) Such backgrounds are dynamically altered as the storyline goes on. The HPD dataset, as the first dataset to facilitate the study of dialogue agent construction for characters within a story, provides rich contextual information about each dialogue session such as scenes, character attributes, and relations. More importantly, all the background information will change over the course of the story. In addition, HPD could support both dialogue generation and retrieval tasks. We evaluate baselines such as DialogGPT and BOB to determine the extent to which they can generate Harry Potter-like responses. The experimental results disappoint us in that although the generated responses are fluent, they still seem out of character for Harry. Besides, we validate the current most robust dialogue agent, ChatGPT, which also can't generate plausible Harry-Potter-like responses in some cases, either. Our results suggest that there is much scope for future research.

1 Introduction

Building intelligent dialogue systems is a long-standing pursuit of the NLP community. Recently, a growing body of research (Miller et al., 2017; Zhang et al., 2018a, 2020; Zhou et al., 2021; Gu et al., 2022; Song et al., 2021; Chen et al., 2021; Thoppilan et al., 2022; You et al., 2021) has achieved great progress in this area. Although previous works have attempted to endow dialogue systems with persona (Zhang et al., 2018a; Song et al., 2021), emotions (Liu et al., 2021), and knowledge (Miller et al., 2017; Thoppilan et al., 2022), this paper takes a further step to build dialogue agents for characters in a story.

Dataset	Sce.	Atr.	Rel.	Dy.	Sl.
Douban (2017)	×	×	×	×	×
Ubuntu (2015)	×	×	×	×	×
TaoBao (2018b)	×	×	×	×	×
PchatbotW (2021)	×	√	×	×	×
PeDialog (2019)	×	√	×	×	×
KvPI (2020)	×	√	×	×	×
P-CHAT (2018a)	×	√	×	×	×
WOW (2019)	√	√	×	×	×
Fri-QA (2019)	×	√	√	×	×
Focus (2021)	√	√	√	×	×
Ours	√	√	√	√	√

Table 1: Datasets Comparison. **Sce.**, **Atr.**, **Rel.**, **Dy.** and **Sl.** denote **Scenes**, **Attributes**, **Relations**, **Dynamic** and **Storyline**, separately.

Building such dialogue systems poses new challenges: First, it is indispensable for deeply understanding the storyline over each dialogue session, which refers to what happened before the conversation. Second, all types of storyline-related information, such as scene, character relations and character attributes, are time-sensitive, they dynamically change as the story goes on. Modeling the storyline and these dynamic background information effectively is crucial for dialogue agents to mimic the character behaviors in a specific conversation.

The above characteristics are not included in existing dialogue datasets. As shown in Table 1, although some related studies (Zhang et al., 2018a; Zheng et al., 2019; Jang et al., 2021) already correspond dialogues to scene, relations, and attributes, one main issue is that such corpora only contain static information without any changes, and their

[*]Work done when interned at Tencent AI Lab. [§] Indicates Corresponding authors.

dialogues sessions do not accordance to any storyline. To this end, we propose Harry Potter Dialogue (HPD), the first dialogue dataset that integrates with **scene**, **attributes** and **relations** which are **dynamically** changed as the **storyline** goes on. Our work can facilitate research to construct more human-like conversational systems in practice. For example, virtual assistants, NPC in games, etc.

Concretely, we collect data from the Harry Potter Series in the hopes of creating a Harry-Potter-like dialogue agent. We hire some professional annotators who are also addicted to Harry Potter fans to extract and annotate the dialogues in the books. Besides collecting dialogue contexts, we also annotate the following rich and fine-grained background information based on the collected dialogues: 1) We first annotate the speakers of each dialogue so that the model can aware of who he is talking with. 2) With the goal of giving a full picture of each character in dialogue, we then annotate three types of background information: scene, relations, and attributes, The **scene** is the paragraphs around a dialogue session, it incorporates information about when, where, and why the dialogue took place. The **relation** is a matrix indicates the fine-grained relations between Harry and other characters. The **attribute** is another matrix that records the main attributes of each character, such as gender, age, spells, and belongings[1]. All these background information will change as the story goes on, which means that each dialogue session corresponds to a unique background. In total, we annotate 113 key characters, each having 12 relations to Harry Potter and 13 unique attributes. With this level of detail in the annotations, the ability to construct a Harry-Potter-like dialogue agent is made feasible.

Finally, we collect 1042 dialogue sessions as our training set and further construct a test set with 178 dialogue sessions to evaluate how similar a dialogue system is to Harry Potter. Each session in our test set consists of at least one positive response and 9 negative responses. Importantly, this test set can facilitate the evaluation of both **generation-based** and **retrieval-based** dialogue systems.

We benchmark several strong dialogue systems on our HPD, including three generation-based models and one retrieval-based model. Automatic and human evaluation results show that generated responses from these models are still far-away from the high-quality Harry Potter-like responses, indicating there is a large headroom for improvement.

[1]Annotation costs mainly come from character attributes and relationships, which are dynamically annotated by each chapter. There are more than 100 important characters in novels. If we annotate all relations between each character will result in great expensive costs in time and money. Therefore, we only consider modeling the conversation of Harry in this work. Nevertheless, we also collect all character attributes for potential use in modeling conversations of other characters.

2 Task Definition

We use the Harry Potter Series as our test-bed with the aim of creating dialogue agents for characters in a story. The generated responses of such this dialogue agent should be not only relevant to the context, but also seems like something Harry would say at this time and scene.

Figure 1 shows some main factors that affect behaviors of Harry in a conversation. The first factor is the conversation history, which is the most important factor that determines Harry's response. The scene, which is the second factor, provides details about the motivation (*The Hippogriff Buckbeak attacked Draco Malfoy at Hagrid's first Conservation of Magical Creature lesson*) of this dialogue. The third factor is the participants' information (**attributes and relations**), obviously Harry's behaviors will be totally different from Hermione and Malfoy. The latter two factors are exactly determined by the timing of this dialogue (*Book 3-Chapter 5*), so they are continuously varied over the storyline.

Formally, the task of building dialogue agents for characters in a storyline can be defined as follows: Given a dialogue history $\mathbf{H}$, corresponding dialogue scene $\mathbf{S}$ and participants information $\mathbf{P}$ as input, which are changed depending on the development of storyline. The dialogue agent is supposed to generate a response $\mathbf{Y} = y_1, y_2, ..., y_n$:

$$\mathcal{Y} = \underset{Y}{\operatorname{argmax}} P(\mathbf{Y}|\mathbf{H}, \mathbf{S}, \mathbf{P})$$

It is worth noting that the format of $\mathbf{H}$, $\mathbf{S}$, and $\mathbf{P}$ are not strictly limited. In this paper, we will format $\mathbf{H}$ and $\mathbf{S}$ as natural language sentences, and provide $\mathbf{P}$ in key-value pair formats. $\mathbf{Y}$ is supposed to be not only fluent and natural, but also highly relevant to $\mathbf{S}$ and $\mathbf{P}$.

3 Dataset Construction

A high-quality dataset including all pertinent information in section 2 is the prerequisite for building dialogue agents for characters in a story. Unfortunately, so far there are currently no publicly

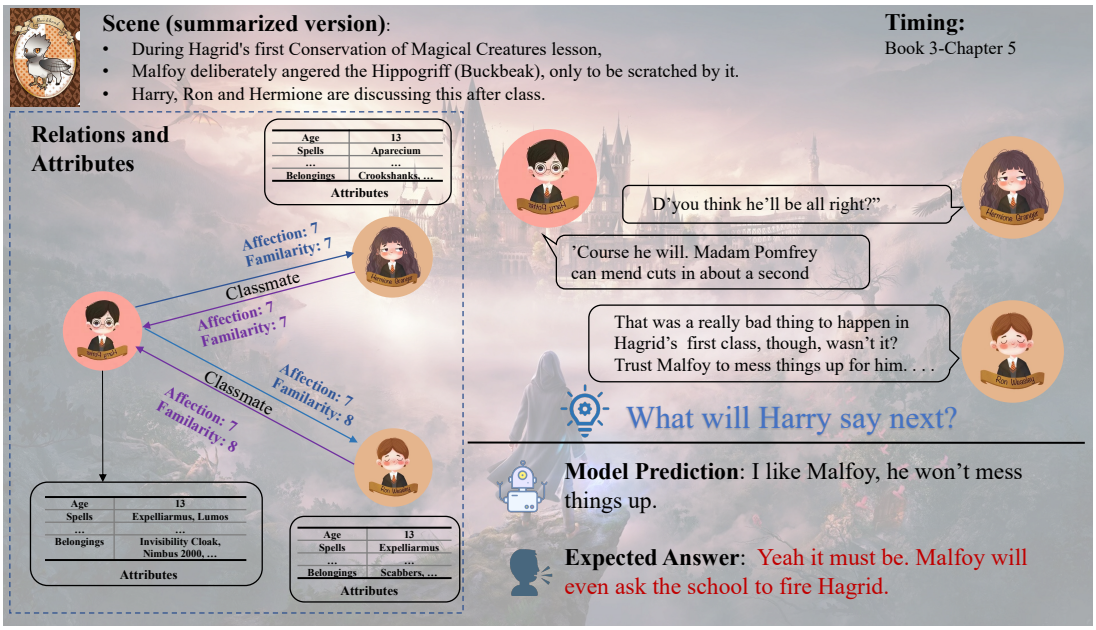

Figure 1: A conversation example selected from Book 3-Chapter 5 in Harry Potter Series, which involves Harry, Hermione and Ron. In this example, we present scene and timing of the conversation, and relations and attributes of speakers. Texts in red refers to the expected response.

available datasets that provide information about the dialogue scene and participants. To facilitate the study of this task, we construct a new dataset from the popular fictions Harry Potter Series, in the hopes of creating a Harry-Potter-like dialogue system. All dialogue sessions that harry participated in are collected in this dataset, along with fine-grained annotated dialogue scenes and participant information. The annotation work in this study is done by four addicted fans of Harry Potter.

We collect the following parts to construct our dataset, as shown in Figure 1: The dialogue part (Section 3.1) contains all utterances in the dialogue sessions, as well as the speaker's name of each utterance. The scene part (Section 3.2) includes the text around the dialogue session, whose length ranges from several paragraphs to a whole chapter. Finally, the speaker information part (Section 3.3), which consists of attributes and relations of characters, is shown in the left part of Figure 1. Please note that these scenes, attributes, and relations are time-sensitive, they may change as the story plots go on, which are shown with color words in the selected example of Figure 2. So the annotators are requested to annotate them session by session.

3.1 Dialogue Construction

Dialogue sessions in the books have been divided into two folds: a training set and test set. Their main difference is that each training session contains 1 positive response only, while each test dialogue session consists of at least 1 positive response and 9 negative responses.

Training Set we request the annotators to extract all multi-turn dialogues from the books. Besides, the speaker name of each utterance in the session is labeled as well.

Test Set The effectiveness of dialogue models should be evaluated using a well-designed test set. However, if we directly select the test dialogue sessions from the books, it may meet serious dialogue leakage problems, i.e. the fact that we evaluate on data that were also present in the pre-training corpus. To prevent this problem, we deliberately design a test set in the following steps:

- First, we manually select some raw dialogues which meet the following requirements: (i) Dialogues with only one speaker, which contains only one or two sentences; (ii) Dialogues, in which no response from other speakers of the last question. For these samples, we pick out dialogues that are relevant to Harry and can be answered from Harry's perspective to construct the test set.
- Second, directly asking the annotators to compose an acceptable and natural response from scratch for each dialogue is quite challenging. We thus fine-tune several pre-trained dialogues models on our training set, and use them to generate potential responses for each selected dialogue session. In detail, we fine-tune two types of models: generation-based

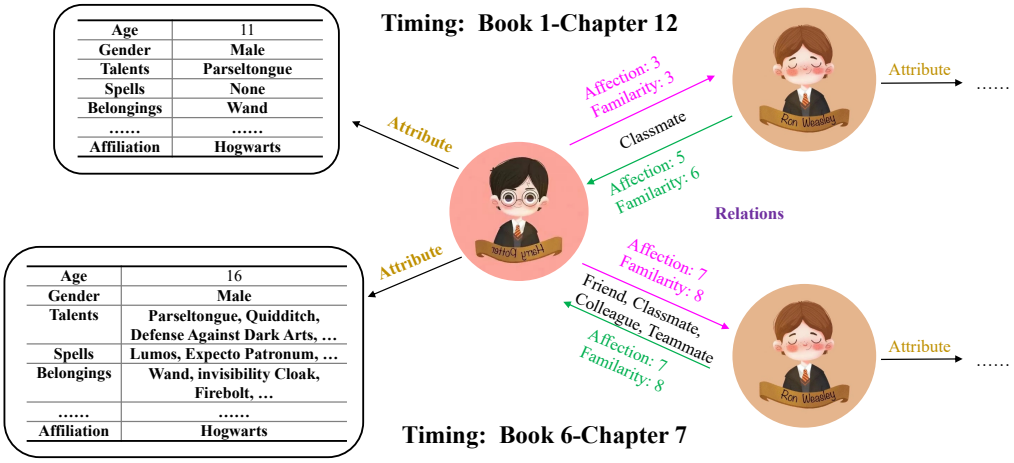

Figure 2: Data examples from two different timings: Book 1-Chapter 12 and Book 6-Chapter 7. Words in red denote the changed relations and attributes.

and retrieval-based dialogue models in Harry Potter novels. Subsequently, each model is required to predict 10 responses as the candidates of the test set.

- Third, we ask three annotators to select the most reasonable responses from the predictions as the positive response (ground-truth)[2] while others are regarded as negatives. And the other one is responsible for integrating their annotations. If all predictions are considered unreasonable texts, the annotator would write a reasonable response from scratch.

This setup has advantages over employing a single annotator to label predictions: One annotator may regard some predictions as positive responses while others may label them as negatives. Hence, the other annotator can be used to measure the quality of annotated answers when they have disagreements. These strategies help to prevent spam and bias, and thus get high-agreement data. Finally, we manually double-check and revise mistakes to further ensure the data quality.

3.2 Scenes Construction

In order to offer accurate location information and textual details for each dialogue, we further annotate scenes. In our settings, the text that surrounds and immediately relates to the dialogue is regarded as the scene. Annotators are required to label the beginning and end of a scene in units of paragraphs. In other words, scenes are composed of several consistent paragraphs. There are two guiding principles: The first is that the longest scene cannot be longer than one chapter, and the second is that no scene can contain more than one dialogue session.

Similarly, three of the annotators are responsible for scene annotation, and the remaining one is responsible for integrating their annotations to obtain high-agreement data.

3.3 Attributes and Relations Construction

One of the most important and appealing properties of our benchmark is fine-grained annotated character information, which includes the attributes and relations of the characters. With the goal of providing in-depth and comprehensive character information, we collect 13 fine-grained attributes and 12 relations.

We divide the attributes into two categories: (1) inborn; (2) nurture. The former denotes some innate attributes or abilities, which contains `Gender`, `Age`, `Lineage`, `Talents` and `Looks`. The latter refers to properties through acquired efforts or opportunities, including `Achievement`, `Title`, `Belongings`, `Export`, `Hobby`, `Character`, `Spells` and `Nickname`, as some cases are presented in Figure 2. In total, we collect 13 attributes for each character, which basically covers most properties in Harry Potter series.

The relations between Harry and other characters can be classified into binary relations and discrete relations. The former includes 8 types, which are `Friend`, `Classmate`, `Teacher`, `Family`, `Lover`, `Opponent`, `Teammate`, `Enemy`. Multiple binary relations can exist between two characters. Harry and Ron, for instance, are friends, classmates, and teammates in the Quidditch team (In Book-6 only).

[2]Some questions may have multiple valid answers

Affection	
10	Parents and relatives who died for Harry. Examples: Harry's parents
9	People who very close with Harry and save Harry's life. Examples: Ron in Book-7
8	People who in love with Harry. Examples: Ginny
7	Best friends Examples: Ron in Book-2
6	Close Friends and very kindly to Harry. Examples: Hagrid
5	People who Often helps Harry. Examples: Dumbledore
4	Familiar Classmates/Teachers and people who are relatively friendly to Harry. Examples: Neville
3	Normal Teammates. Examples: Wood
2	Normal Classmates/Teachers. Examples: Lavender Brown
1	First Metting. Examples: Harry first met Ron and Hermione
0	Stranger
-2	Rude/Frivolous/Rude/Very strict and mean teachers or classmates. Examples: Draco Malfoy in Book-1, Filch
-4	Deliberately bullying/deliberately targeting. Examples: Snape, Dudley
-6	Maliciously targeting and harm. Examples: Draco Malfoy in Book-5
-8	Intentionally inflict harm. Examples: Bellatrix Lestrange in Book-5
-10	Kill Harry's parents. Examples: Voldemort

Figure 3: `Affection` Rules and Examples.

We take into account the fact that each character's level of familiarity and affection with Harry might vary and is always shifting as the storyline goes on. For instance, Harry has a far higher level of familiarity and affection with Ron than he does with Snape. Therefore we collect 4 types of discrete relations: (1) Harry's `Familiarity` to someone, (2) Harry's `Affection` to someone, (3) someone's `Familiarity` to Harry, and (4) someone's `Affection` to Harry. Two examples can illustrate the difference between them: Draco Malfoy hates Harry, but he is also familiar with Harry's habits, so his Affection to Harry is low but his familiarity to Harry is high. Moreover, since Harry lost his parents, Dumbledore has shown a lot of concern for Harry, so his familiarity and affection to Harry are high, but Harry's familiarity to Dumbledore is low. Another example of how attributes and relations changed over storyline is shown in Figure 2. In Book 1-Chapter 7, Harry just meets Ron at the train to Hogwarts, and now he is a stranger to the wizarding world. So his affection and familiarity to Ron is relatively low (1 and 2, respectively), and he doesn't aware of any spells. However, when story goes on, at Book 6-Chapter 7, he is a full-fledged wizard, and Ron is his best friend. So their affection and familiarity are high at this time (7 and 8 respectively). Harry also masters a lot of spells such as *Expecto Patronum* and *Expelliarmus*, and has some powerful equipment such as his broomstick *Firebolt* and the *invisibility cloak*.

`Affection` Annotation `Affection` is rated on a twenty-level, ranging from -10 to 10, where -10 and 10 indicate the lowest and highest affection, separately. And `Score(Affection)>0` denotes the character has the positive relationship with Harry whereas `Score(Affection)<0` means the character has the negative relationship with Harry.

Familiarity	
10	Close friends stay together for years and are very familiar with each other's habits, secrets and temper.
	Examples: Ron in Book-7
8	Stay together and are familiar with each other.
	Examples: Vernon in Book-1
6	Characters who know and are familiar with each other's background information.
	Examples: Sirius in Book-5
4	Friends/Teachers meet multiple times and are slightly familiar with each other.
	Examples: Professor. McGonagall
2	Familiar with the character's background but don't know him/her.
	Examples: Ron before met Harry in Book-1
1	Meet for the first time
0	Stranger

Figure 4: `Familiarity` Rules and Examples.

Prior to going further, we first define the rules of annotating `Affection`, which are shown in Figure 3. During annotation, we also provide comprehensive descriptions and examples for each `Affection` level to help workers. Concretely, `Score(Affection)=1` refers to *someone meeting Harry for the first time*. For instance, when Hary first met Ron and Hermione in Book 1, their' `Affection` to Harry and Harry's `Affection` to them are both set to 1. And `Score(Affection)=-10` means *someone killed Harry's parents*, where Voldemort meets this condition in the novels.

`Familiarity` Annotation Similarly, we also rate `Familiarity` with 10 level, which ranges from 0 to 10, where 10 indicates the highest affection and 0 indicates the lowest affection. We present detailed explanations for each `Familiarity` level in Figure 4. Concretely, `Score(Familiarity)=0` denotes stranger, and `Score(Familiarity)=10` denotes close friends who often stay together for many years and are very familiar with each other's habits, secrets and temperaments, where Ron meets this condition in Book 7.

During annotation, we ask each annotator following these rules to annotate `Familiarity` and `Affection`. In order to speed up the annotation speed and quality, we pre-annotated some examples to guide the annotators. Considering `Affection` and `Familiarity` work both ways, three annotators are required to annotate someone's `Affection`/`Familiarity` to Harry and Harry's `Affection`/`Familiarity`

Statistics	Train	Test
per dialogue		
Average Turns	13.8	8.6
Maximum Speakers	20	8
Minimum Speakers	2	2
per sentence		
Average Length	32.9	28.3
Maximum Length	77	31
Minimum Length	3	7
Total Dialogues	1042	178

Table 2: Data statics of collected dialogues.

to someone chapter by chapter. And another annotator is responsible for measuring the quality of annotated data when the other three annotators have disagreements. To further control the data quality, we manually re-check and revise some controversial annotations.

Claim Notice that we hope to provide as rich character information as possible for the community, even if they seem redundant in this work. Therefore, we collect 13 attributes for each character and 12 relations in the collected HPD. We leave plenty of opportunity for other research communities to investigate which information is helpful in their work. They also can build other tasks such sentiment analysis of Harry Potter based on our fine-grained annotations. In other words, it is not required to include all fine-grained annotated information in the study. For instance, we only try to utilize `Familiarity` and `Affection` information in our experiments while other annotated knowledge are left for future. We believe redundant is better than lack because researchers can determine which information to use by themselves.

3.4 Data Statistics

The detailed statistics of dialogues are shown in Table 2, training set and test set contain 1042 and 178 dialogues, separately. It is worthwhile to notice that, we initially collect 1471 dialogues for constructing the training set, and we filter out those dialogues that are without Harry[3], leading 1042 conversations for consideration. Obviously, we also can observe the maximum speaker per dialogue reaches 20, indicating our dataset is very challenging and complicated. Considering not all characters are essential to understanding and driving the story in Harry Potter series, we choose 113 important characters to annotate their attributes and relations, such as Harry, Ron and etc.

[3]Considering that some of them are important for understanding the storyline, we've kept them for future work.

4 Baselines

To investigate how similar existing dialogue systems can be to Harry Potter, we build four baselines in our experiments. They can be divided into two types: generation-based systems and retrieval-based systems.

Generation Models We choose three popular pre-trained generation models: GPT-2 (Radford et al., 2019), Dialog-GPT (Gu et al., 2022) and Bert-over-Bert (BOB) (Song et al., 2021) as our backbones, which have demonstrated strong performances on the task of open-domain dialogue (Zheng et al., 2019; Dinan et al., 2019). In order to validate the effect of annotated persona knowledge, we train BOB with two different **input settings**: (1) BOB only takes each dialogue session as input, called *Ori-BOB*; (2) BOB takes both each dialogue session and speaker personas as input, which includes character relations, named *Per-BOB*. Concretely, we made a preliminary attempt to transfer annotated `Familarity` and `Affection` into natural language text. For instance, Harry's `Familarity` with Hagrid is 5, which can be explained as "*Harry is relatively familiar with Hagrid.*". In this way, we hope the model will learn to capture character relationships.

Retrieval-based Model Retrieval-based models are required to select the best response to a given dialogue context from a set of candidate responses. Since we already provide at least 9 negative samples for each dialogue session in the test set, the task of our retrieval-based model is to select the positive responses from all possible candidates. In this paper, we select BERT-FP (Han et al., 2021) as our baseline model. More details about training these baselines can be seen in Appendix B.

5 Experiments

5.1 Evaluation Metric

We evaluate the response from baseline models from two main aspects: *response quality* and *persona consistency* of Harry. Given that human judgment is the most thorough and realistic assessment of whether the generated text is Harry-Potter-like in our task. We recruit 4 annotators to evaluate the quality and consistency of generated responses

Model	Flue.	Relv.Sce.	Relv.Att.	Relv.Re.
GPT-2	3.34	1.68	1.9	1.5
Dialog-GPT	**4.27**	**2.38**	**2.5**	2.04
Ori-BOB	4.01	1.93	1.78	1.9
Per-BOB	4.2	2.01	2.16	**2.15**

Table 3: Human evaluation results of three generation models on the testset of HPD.

Model	Flue.	Relv.Sce.	Relv.Att.	Relv.Re.
ChatGPT	49	43	40	31

Table 4: Human evaluation of ChatGPT.

on HPD. For a fair comparison, we also employ **automatic evaluations**, seen in Appendix C.

Human Metrics All annotators are asked to evaluate the persona consistency from four criteria: Fluency (**Flue.**), Relevance with the Scene (**Relv.Sce.**), Relevance with the Attributes (**Relv.Att.**) and Relevance with the Relations (**Relv.Re.**). Concretely, the latter three evaluate the generated response in line with Harry Potter. Each criterion is rated on a five-scale, where 1, 3 and 5 represent unreasonable, moderate and perfect performance, respectively. Rating explanations can be found in Appendix C.1.

5.2 Results

Human Evaluation The human evaluation results are shown in Table 3. From the table, we can observe the following conclusions: (1) All generation models performed well on **Flue.**, where GPT-2, Ori-BOB, Per-BOB and Dialog-GPT obtain 3.34, 4.01, 4.2 and 4.27 scores separately, which are consistent with their performances in automatic evaluation results. (2) All models achieve good performances on **Flue.**, but no one outperforms 3 scores on each **Relv.Sce.**, **Relv.Att.** and **Relv.Re.**. The results indicate although their generated responses seem natural but can not be seen as qualified Harry Potter-like responses, which don't capture the attribute and relation information within the given scene, or even violate these properties. (3) Per-BOB shows better results than Ori-BOB, especially on **Relv.Re.**, proving incorporating `Familarity` and `Affection` information into the input indeed help the model understand the relations between speakers. Performances of these models in terms of human evaluation basically match their results on Δ**M** and Δ**P** (Table 6).

Besides, we manually validate whether ChatGPT can generate responses that are consistent with Harry Potter using the above human evaluation perspectives. Concretely, we randomly sample 50 examples from our test set and report the number of generated responses consistent with Harry Potter in each view in Table 4. We can find almost all responses are fluent, and about 90% are relevant with the scene. However, ChatGPT also doesn't capture speaker relations under about 40% cases.

Generally, all results prove how to effectively incorporate character relations and attributes knowledge over each dialogue session is still to explore, leaving much room for communities.

6 Case Study

In this section, we present some sampled cases in Table 5, and analyze some typical errors of these baseline models. We show more case studies in Appendix D (Table 7).

Relationship Contradiction Seen in Case 1 from Table 5, we can find the scene of this dialogue can be summarized as *"Dumbledore is comforting Harry to get out of the shadow of Sirius' death."* The given relations denotes Harry has high affection and familiarity towards Dumbledore and Sirius. Therefore, the expected response from Harry should be polite and friendly at least. Most likely he will follow Dumbledore's advice. Nevertheless, generated responses from baselines are universal but far from meeting our expectations. Concretely, responses from GPT and two BOB models are beyond the relationship between Harry and Dumbledore. Moreover, the response from Dialog-GPT is against the close relationship between Harry and Sirius. ChatGPT has a comprehensive and accurate understanding of Harry Potter story, and generates the response that is consistent with the Harry's relation to Dumbledore but against with the scene in this case. In other words, the topic of this dialogue context is not about *"prophecy"*.

Attributes Contradiction Seen in Case 2 from Table 5, The scene of selected dialogue is about *"Harry finds himself able to converse with snakes."* and Harry is facing questioning from Pears, who often bullies Harry with Dudley. At that time, Harry doesn't know anything about the wizarding world, and hasn't yet known and gone to Hogwarts. However, all generated responses violate this fact, which involved some characters or things related to Hogwarts. For instance, the generated response from BOB even states that he has an owl. But in the

Case-1	Case-2
Scene "And now, Harry, on a closely related subject . . . As you so rightly said, Sirius would not have wanted you to shut yourself away."	**Scene** After lunch they went to the reptile house. . . But worst of all, for Harry at least, was Piers calming Belongings enough to say, "Harry was talking to it, weren't you, Harry?"
Timeline–Book 6-Chapter 4 **Speakers**–Harry and Dumbledore	**Timeline**–Book 1-Chapter 2 **Speakers**–Harry, Snake, Piers and Dudley
Harry Attributes Age: 16 Talent: Quidditch, Parseltongue ... Affiliation: Hogwarts Belongings: ... Firebolt-Broom, Marauder's Map	**Harry Attributes** Age: 11 Talent: Parseltongue ... Affiliation: None Belongings: None
Relations Harry's affection for Dumbledore : 6 Harry's familiarity with Dumbledore : 6 Harry's affection for Sirius : 8 Harry's familiarity with Sirius : 6 ...	**Relations** Harry's affection for Dudley : -4 Harry's familiarity with Dudley : 6 Harry's affection for Piers : -3 Harry's affection for Piers : 2 ...
Dialogue ... Dumbledore: Although I think you ought to relax it in favor of your friends, Mr. Ronald Weasley and Miss Hermione Granger. Yes, I think they ought to know. You do them a disservice by not confiding something this important to them. Harry: I didn't want — Dumbledore: — to worry or frighten them? Or perhaps, to confess that you yourself are worried and frightened? You need your friends, Harry. As you so rightly said, Sirius would not have wanted you to shut yourself away.	**Dialogue** ... Dudley: Make it move. Do it again...This is boring. Snake: I get that all the time. Harry: I know, it mush be really annoying. Where do you come from, anyway? Was it nice there? Oh, I see — so you've never been to Brazil? Snake: Brazil, here I come. . . . Thanksss, amigo. Dudley: Out of the way, you! Pears: Harry was talking to it, weren't you?
Generated Responses GPT: I do not need friends. Dialog-GPT: I know him, he wants me shut myself away. Ori-BOB: I think they are meaningless... Per-BOB: It is not your business, it's my own choice. ChatGPT: Okay, I will tell them about the prophecy.	**Generated Responses** GPT: I do not know... I want to go to Hogwarts. Dialog-GPT: Sorry, Professor Snape, this is an accident. Ori-BOB: I think so, Hagrid. Per-BOB: Let us go to see Hagrid with my owl. ChatGPT: I was just asking where it came from.
Expected Response **Harry: Yes, Sirius really doesn't want that, maybe I should talk to them.**	**Expected Response** **Harry: I don't know, you stop talking nonsense, please stop targeting me.**

Table 5: Data examples of HPD and generated responses by baseline. Simply, we only present Harry's attributes and discrete relations to other speakers in the example.

novel's timing (Book 1-Chapter 2), he didn't know Hagrid under the circumstances, let alone an owl. In this scene, Harry is likely trying to downplay the situation, keep his secret and avoid potential conflict with Piers and Dudley, and thus, he unlikely responded with *"I was just asking where it came from..."* that generated from ChatGPT.

In summary, the error analysis shows the existing approaches fall well short of generating smooth and reasonable Harry-like responses. Even incorporating some character relation information into Per-BOB, the generated responses appear slightly reasonable in only a few instances (Table 7). Moreover, ChatGPT, as the current most strong dialogue agent, still doesn't address the challenge in HPD in some situations.

7 Conclusion

In this paper, we propose a new benchmark named Harry Potter Dataset (HPD) to promote building dialogue agents for characters in story. Unlike existing datasets, HPD not only contains interesting dialogues, but also consists of scenes, character attributes and relations which are dynamically changed as the storyline goes on. It also provides a well-designed test set to facilitate the evaluation of both generation-based and retrieval-based dialogue agents. Concretely, annotated scenes are several paragraphs which are around a dialogue session. Considering expensive manually-labeled costs, attributes and relations of each character are provided as key-value pairs. Results and case studies prove performances of current state-of-the-art models on HPD are far away from human expectations, prov-

ing there is ample room for improvement. We hope this work can draw more attention to efficient person-like dialogue modeling in the virtual world.

Ethical Statement

To avoid the potential issue of using Harry Potter Novels, we promise the annotated dataset is developed for non-commercial use. Moreover, we only provide the line number and page number of each collected dialogue in Harry Potter novel rather than give the detailed content of each dialogue session. We further supply the script to extract corresponding raw dialogue data from the novels according to the provided line and page numbers, in which the data format is the same as the data examples in Table 5 and Table 7. As for the annotated character attributes and relations, we have our own copyright and will release for research communities.

Limitations

The main target of this paper is towards building dialogue agents for characters in a story. In this paper, we present a new benchmark named Harry Potter Dialogue (HPD) in the hope of creating a Harry-Potter-like dialogue agent. The significant feature of HPD is that it contains detailed scenes and fine-grained attributes and relations of each speaker which are dynamically changed as the storyline goes on. More generally, we expect the core idea of this paper can give insights into other research communities that want to build effective person-like chatbots in the virtual world. Our fine-grained annotated knowledge also can be used to build other tasks such as sentiment analysis and reading comprehension of Harry Potter. Admittedly, the data of the proposed dataset from Harry Potter Series are restricted to the specific area, that is, Harry Potter Magic World. Considering the high annotation cost, our character relation annotation works are restricted to Harry Potter. These concerns warrant further research and consideration when utilizing this work to build intelligent person-like dialogue systems in the virtual world.

References

Nuo Chen, Chenyu You, and Yuexian Zou. 2021. Self-supervised dialogue learning for spoken conversational question answering. In *Interspeech*, pages 231–235. ISCA.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *CMCL@ACL*, pages 76–87. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *ICLR (Poster)*. OpenReview.net.

Yuxian Gu, Jiaxin Wen, Hao Sun, Yi Song, Pei Ke, Chujie Zheng, Zheng Zhang, Jianzhu Yao, Xiaoyan Zhu, Jie Tang, and Minlie Huang. 2022. EVA2.0: Investigating open-domain chinese dialogue systems with large-scale pre-training. *arXiv preprint arXiv:2203.09313*.

Janghoon Han, Taesuk Hong, Byoungjae Kim, Youngjoong Ko, and Jungyun Seo. 2021. Fine-grained post-training for improving retrieval-based dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1549–1558, Online. Association for Computational Linguistics.

Yoonna Jang, Jungwoo Lim, Yuna Hur, Dongsuk Oh, Suhyune Son, Yeonsoo Lee, Dong-Hoon Shin, Seungryong Kim, and Heuiseok Lim. 2021. Call for customized conversation: Customized conversation grounding persona and knowledge. *CoRR*, abs/2112.08619.

Satwik Kottur, Xiaoyu Wang, and Vítor Carvalho. 2017. Exploring personalized neural conversational models. In *IJCAI*, pages 3728–3734. ijcai.org.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and William B. Dolan. 2016b. A persona-based neural conversation model. In *ACL (1)*. The Association for Computer Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *ACL/IJCNLP (1)*, pages 3469–3483. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *SIGDIAL Conference*, pages 285–294. The Association for Computer Linguistics.

A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. 2017. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaïd Harchaoui. 2021. MAUVE: measuring the gap between neural text and human text using divergence frontiers. In *NeurIPS*, pages 4816–4828.

Hongjin Qian, Xiaohe Li, Hanxun Zhong, Yu Guo, Yueyuan Ma, Yutao Zhu, Zhanliang Liu, Zhicheng Dou, and Ji-Rong Wen. 2021. Pchatbot: A large-scale dataset for personalized chatbot. In *SIGIR*, pages 2470–2477. ACM.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. 2021. BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–177, Online. Association for Computational Linguistics.

Haoyu Song, Yan Wang, Weinan 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)*, pages 6651–6662.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, abs/1901.08149.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *ACL (1)*, pages 496–505. Association for Computational Linguistics.

Zhengzhe Yang and Jinho D. Choi. 2019. Friendsqa: Open-domain question answering on TV show transcripts. In *SIGdial*, pages 188–197. Association for Computational Linguistics.

Chenyu You, Nuo Chen, Fenglin Liu, Dongchao Yang, and Yuexian Zou. 2020. Towards data distillation for end-to-end spoken conversational question answering. *CoRR*, abs/2010.08923.

Chenyu You, Nuo Chen, and Yuexian Zou. 2021. Self-supervised contrastive cross-modality representation learning for spoken question answering. In *EMNLP (Findings)*, pages 28–39. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *ACL (1)*, pages 2204–2213. Association for Computational Linguistics.

Wei-Nan Zhang, Qingfu Zhu, Yifa Wang, Yanyan Zhao, and Ting Liu. 2019. Neural personalized response generation as domain adaptation. *World Wide Web*, 22(4):1427–1446.

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 *ACL, system demonstration*.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018b. Modeling multi-turn conversation with deep utterance aggregation. In *COLING*, pages 3740–3752. Association for Computational Linguistics.

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

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *AAAI*, pages 9693–9700. AAAI Press.

Hao Zhou, Pei Ke, Zheng Zhang, Yuxian Gu, Yinhe Zheng, Chujie Zheng, Yida Wang, Chen Henry Wu, Hao Sun, Xiaocong Yang, Bosi Wen, Xiaoyan Zhu, Minlie Huang, and Jie Tang. 2021. EVA: An open-domain chinese dialogue system with large-scale generative pre-training. *arXiv preprint arXiv:2108.01547*.

A Related Work

Recently, building personalized dialogue systems draw a lot of attention from research communities. Aiming for promoting this area, several efforts and benchmarks (Danescu-Niculescu-Mizil and Lee, 2011; Zheng et al., 2019; Yang and Choi, 2019; Song et al., 2020; Zheng et al., 2020; You et al., 2020) have been made, demonstrating promising results for endowing personal style into dialogue systems. Some initial efforts (Danescu-Niculescu-Mizil and Lee, 2011) aimed modeling characters from movie.

Further developments provide personas via two types: implicit and explicit personalization. In the former stream (Kottur et al., 2017; Li et al., 2016b; Zhang et al., 2019), each speaker personality information can be compressed as the persona embeddings. In this manner, the existing issue of these methods is hard to explain their effectiveness. For the latter (Wolf et al., 2019; Song et al., 2020), the personal information are provided as: (1) *dense personas*, such as speaker profile or text-described personas; (2) *sparse personas*, including some personality traits. For example, personas from (Zheng et al., 2019) are formulated as key-value pairs: "Age:xx, Gender:xx, Location: xx".

More recently, several efforts (Dinan et al., 2019; Yang and Choi, 2019; Jang et al., 2021) incorporated scenes and relations knowledge into each dialogue session for encouraging more real personalized conversation. (Yang and Choi, 2019) presented a open-domain question answering dataset excerpted from *Friends Series*, where each dialogue involves multiple speakers and their relations. (Jang et al., 2021) proposed FoCus dataset where the customized responses are generated based on the user's persona and Wikipedia background knowledge.

In this paper, our goal is to build dialogue agents for the character in a story, which requires modeling scenes and speaker information that are dynamically changed as the storyline goes on. However, the personality settings of current studies are static, and are not changed with scenes or times changing. Therefore, we present HPD: Harry Potter Dialogue Dataset, aiming for creating Harry-Potter-like dialogue agent. In detail, we annotate detailed scenes, attributes and relations of each speaker over given dialogues to help the model have a deeper understanding of the dialogue background information.

B Baselines Setup

In the following, we briefly introduce each model and describe the training and test details.

B.1 Baselines

GPT is a well-known pre-trained language model, which is designed for text generation. We directly fine-tune GPT in Harry Potter Novels with its raw pre-training tasks. During inference, we take the scenes as the model input, along with a prompt "Harry says:" at the end of the scenes.

Dialog-GPT is an open-domain generation dialogue system, which achieves state-of-the-art performances in multiple dialogue datasets. When fine-tuning Dialog-GPT[4] on HPD, for each dialogue session with n utterances, previous n-1 dialogue history utterances are seen as $\mathbf{X}$ and fed into the encoder, and thus, the model is trained to generate the last n-th utterance from the decoder.

BOB is designed for personalized dialogue system, which consists of one BERT-based encoder and two BERT-based decoders[5]. In order to validate the effect of annotated persona knowledge, we train BOB with two different **input settings**: (1) BOB only takes each dialogue session as input, called *Ori-BOB*; (2) BOB takes both each dialogue session and speaker personas as input, which includes character relations, named *Per-BOB*. Concretely, we made a preliminary attempt to transfer annotated `Familarity` and `Affection` into natural language text via some pre-defined rules in Appendix **??**. For instance, Harry's `Familarity` with Hagrid is 5, which can be explained as "*Harry is relatively familiar with Hagrid.*". In this way, we hope model learns to capture the relations between characters.

BERT-FP is a commonly-used strong retrieval-based dialogue system, which devises several post-training objectives. When fine-tuning BERT-FP, given n-1 utterances in each dialogue session, the model is required to find the ground-truth response from candidate answers. Concretely, we first post-train BERT-FP in Harry Potter novels and then fine-tune the resulting model in the collected HPD.

B.2 Experimental Setup

During training, we follow the most original experimental settings (e.g., learning rate and batch

[4]https://github.com/thu-coai/Dialog-GPT/

[5]https://github.com/songhaoyu/BoB

size) of baselines from their initialized works[6]. In detail, we employ large and xlarge version of GPT-2 and Dialog-GPT, separately. For all generation models, the maximum lengths of the encoder and decoder are set to 128 and 64. We fine-tune GPT2-large with 11 epochs in our dataset. When training Dialog-GPT and BOB, the epochs are 10 and 80. Notice that, we initialize BOB from the checkpoint trained on PersonalDialogue (Zheng et al., 2019). When training BERT-FP, the max sequence length and epochs are set to 256 and 20 with 8e-6 learning rate.

C Evaluation Metrics

Automatic Metrics Following the line of previous works, we employ some widely-used metrics to validate the performances of these generative models: **Dist.1** (Li et al., 2016a), **Perplexity** (PPL), **MAUVE** (Pillutla et al., 2021) and Δ**P** (Song et al., 2021). Notice that, Larger Δ**P** denotes the model performs better in distinguishing positives from negatives.

Dist.1: the most commonly-used evaluation metric that evaluate the diversity and informativeness of responses.

PPL: another popular metric that evaluates the fluency and relevance (to context) of responses. Lower perplexity represents better language modeling performance.

ΔP: Δ Perplexity (Δ**P**) is a new metric proposed by (Song et al., 2021). It measures consistency from generation-based models' internal distributions. In order to evaluate models fairly in our settings, we reformulate Δ**P** as :

$$\Delta\mathbf{P} = \frac{\mathbf{PPL.}(Negatives) - \mathbf{PPL.}(Positive)}{\mathbf{PPL.}(Negatives) + \mathbf{PPL.}(Positive)}$$

In this way, we normalize the value of Δ**P** from 0 to 1. Larger Δ**P** denotes model performs better in distinguishing positives from negatives.

MAUVE (Pillutla et al., 2021): an effective metric for evaluating generated text, which measures the token representation distribution closeness between the generated text and human-written text. Higher MAUVE means the generated text more human-likes.

[6]The dataset is in both Chinese and English. The experiments are conducted on the Chinese version.

Generation-based					
Model	**Dist.1**(↑)	**P**(↓)	Δ**P**(↑)	**M**(↑)	Δ**M**(↑)
GPT-2	9.86	18.3	0.12	0.809	-0.164
Ori-BOB	12.87	3.03	0.21	0.940	-0.011
Per-BOB	13.25	2.41	0.08	0.948	0.003
EVA	23.01	37.8	0.54	0.968	0.192
Retrieval-based					
Model	**MAP**	**MRR**	**P@1**	**R10@1**	**R10@5**
BERT-FP	0.468	0.468	0.259	0.259	0.788

Table 6: Automatic evaluation results of three generation-based and one retrieval-based models. **P** and **M** represent **PPL.** and **MAUVE**, separately. Ori-BOB denotes the input of BOB only consists of dialogue sessions. Per-BOB refers to the input of BOB contains dialogue sessions and speaker relations.

Moreover, we propose new metric named Δ**M**, which is formulated as:

$$\Delta\mathbf{M} = \mathbf{MAUVE}(Positive, Response) - \mathbf{MAUVE}(Negative, Response)$$

where a larger Δ**M** denotes the generated text is more Harry-like. Considering negatives in our test set may betray the background of the dialogue or personal information of the speakers, and even has lower fluency. Therefore, larger Δ**M** also represents higher consistency to the scene, attributes, and relations to some extent. In detail, we utilize all collected negatives from each dialogue session in the test set to compute Δ**M** and Δ**P**, and report average results.

For evaluating the retrieval-based model, we also employ some common metrics: MAP (mean average precision), MRR (mean reciprocal rank), and P@1 (precision at one). Recall also be considered, which is used as R10@k, which implies that the correct response exists among the top k candidates out of the ten candidate responses.

Automatic Evaluation We first report four model results on automatic metrics in Table 6: (1) At the first glance of the table, three-generation models achieve relatively high scores on **Dist.1**. Especially, Dialog-GPT achieves the best results with 23.01, which denotes its generated responses have higher diversity. (2) Secondly, we can find Per-BOB has the lowest **PPL.** (2.41), indicating it seems has learned a good dialogue language model fitting our dataset, but lowest Δ**P** (0.08) means it still not has a deep understanding of the background information over each dialogue session. (3)

Thirdly, all models achieve high MAUVE scores, showing that their generated responses are close to the human-written texts in terms of token distribution, especially for Dialog-GPT, which obtains 0.968 MAUVE scores. (4) Lastly, GPT and Ori-BOB get -0.164 and -0.011 scores on Δ**M** validate it can not distinguish negatives and ground-truth responses, that is, the generated responses from GPT-2 and Ori-BOB are universal but not personalized. Interestingly, a similar phenomenon occurs in Per-BOB, which obtains only 0.003 MAUVE score. And Dialog-GPT achieves the best Δ**M** score among these models, indicating its ability of capturing scenes and speaker information is stronger than the other two models.

BERT-FP also performs poorly on $P@1$ score (25.9%) and MAP score (46.8%) in the retrieval-based task. The results show the current state-of-the-art retrieval-based model also can not handle the challenge of our benchmark, and thus, there is ample room for future improvement.

C.1 Human Evaluation Rules

For instance, **Flue.** 1 denotes the generated response is cluttered and unreadable and **Flue.** 3 means the generated response is basically readable but has some minor errors. As for latter three metrics, **Relv.Sce./Relv.Att./Relv.Re.** 1 refers generated responses completely violate scenes, Harry's attributes and relations with other characters. **Relv.Sce./Relv.Att./Relv.Re.** 3 denotes generated responses relatively reflect some information about the scenes, Harry's attributes and relations with other characters but still contradict some facts in the specific storyline. **Relv.Sce./Relv.Att./Relv.Re.** 5 indicates generated responses basically conform to the scenes, Harry's attributes and relations with other characters and have no obvious factual errors.

D More Case Study

Dynamic attributes and relations Contradiction Intuitively, a higher pursuit of our task is expect the dialogue system can generate the logical response according to dynamic attributes and relations between characters. We present case studies to verify the ability of baselines. Two cases are shown in the Table 7. In Case 3, the context of scene is *"Ron called Harry, only to be picked up by Vernon. Vernon is unhappy with this and is berating Harry."*. Considering, Ron is one of his best friend (can be inferred from their affection and familiarity) at that time, the expected response from Harry should reflect this. In contrast, in Case 4, the dialogue occurs under the background of *"Harry and Ron misunderstood each other over the Goblet of Fire thing, and they were **very angry** with each other. At this point, Hermione told Harry to go to the Hogsmeade village to relax."*. Therefore, Harry's affection for Ron has a large drop (-12) when changing scene in Case 3 to scene in Case 4. Specifically, Harry must have been reluctant to see Ron at the timing of Case 4, and thus, the expected response should contain the similar meanings. Unfortunately, the generated response from current baselines don't take these dynamic attributes and relations into consideration.

Case 3	Case 4
Scene Harry was particularly keen to avoid trouble with his aunt and uncle at the moment, as they were already in an especially bad mood with him, all because he'd received a telephone call from a fellow wizard one week into the school vacation... And he threw the receiver back onto the telephone as if dropping a poisonous spider. The fight that had followed had been one of the worst ever."HOW DARE YOU GIVE THIS NUMBER TO PEOPLE LIKE — PEOPLE LIKE YOU !" Uncle Vernon had roared, spraying Harry with spit.	**Scene** So Harry put on his Invisibility Cloak in the dormitory, went back downstairs, and together he and Hermione set off for Hogsmeade. Harry felt wonderfully free under the cloak; he watched other students walking past them as they entered the village, most of them sporting Support Cedric Diggory! badges, but no horrible re- marks came his way for a change, and nobody was quoting that stupid article..."Why don't we go and have a butter beer in the Three Broomsticks, it's a bit cold, isn't it? You don't have to talk to Ron!" she added irritably, correctly interpreting his silence.
Timeline Book 3-Chapter 1 **Speakers** Harry, Vernon and Ron	**Timeline** Book 4-Chapter 19 **Speakers** Harry and Hermione
Harry Attributes Gender: Male Age: 13 Talent: Quidditch, Parseltongue ... Affiliation: Hogwarts Lineage: Pure Blood Wizard Own: Owl, Invisibility Cloak, Pocket Looking Glass	**Harry Attributes** Gender: Male Age: 14 Talent: Quidditch, Parseltongue ... Affiliation: Hogwarts Lineage: Pure Blood Wizard Own: Holly Phoenix Feather Wand, Owl, Invisibility Cloak, Pocket Looking Glass, Firebolt-Broom, Marauder's Map
Relations Harry's affection for Vernon : -5 Harry's familiarity with Vernon : 5 Harry's affection for Ron : 7 Harry's familiarity with Ron : 7 **Family** Parents: James Potter and Lily Potter Relatives: Dudley, Aunt Petunia and Uncle Vernon Friends: Ron, Hermione and etc ...	**Relations** Harry's affection for Ron : -5 Harry's familiarity with Ron : 7 Harry's affection for Hermione : 7 Harry's familiarity with Hermione : 7 **Family** Parents: James Potter and Lily Potter Godfather: Sirius Relatives: Dudley, Aunt Petunia and Uncle Vernon Friends: Hermione... ...
Dialogue Vernon: Vernon Dursley speaking. Ron: HELLO? HELLO? CAN YOU HEAR ME? I — WANT — TO — TALK — TO — HARRY — POTTER! Vernon: WHO IS THIS? WHO ARE YOU? Ron: RON — WEASLEY! I'M — A — FRIEND — OF — HARRY'S — FROM — SCHOOL — ... Vernon: HOW DARE YOU GIVE THIS NUMBER TO PEOPLE LIKE — PEOPLE LIKE YOU !	**Dialogue** Hermione: People keep looking at me now. They think I'm talking to myself. Harry: Don't move your lips so much then. Hermione: Come on, please just take off your cloak for a bit, no one's going to bother you here. Harry: Oh yeah? Look behind you. ... Hermione: Why don't we go and have a butterbeer in the Three Broomsticks, it's a bit cold, isn't it? You don't have to talk to Ron!
Generated Responses GPT: He is finding for Sirius. Dialog-GPT: I know him a lot. Ori-BOB: I think so, we can ... Per-BOB: I think I can have a call with... **Expected Response**: It's not my fault, Ron is a good friend of mine and I just want to keep in touch with them when I'm on vacation.	**Generated Responses** GPT: I want to go to see Ron. Dialog-GPT: He is not safe. Ori-BOB: We are the same. Per-BOB: Let us go to see Ron. What is he doing there? **Expected Response**: I'll go, but I don't want to meet Ron, I'm going to put on my invisibility cloak.

Table 7: Data example of our dataset and generated responses by baselines (2).