HOKIEBOT: Towards Personalized Open-domain Chatbot with Long-term Dialogue Management and Customizable Automatic Evaluation

Ying Shen, Jingyuan Qi, Sijia Wang, Barry Menglong Yao,
Minqian Liu, Zhiyang Xu, Trevor Ashby, Lifu Huang
Computer Science Department
Virginia Tech

Abstract

With the increasing prevalence of smart speakers and virtual chatbots, dialogue systems have become a prominent research area in natural language processing. The primary objective of designing a dialogue agent is to converse with humans on a wide range of requests, ranging from general chat to information seeking, event discussion, and more. While existing dialog systems still face numerous challenges, we mainly focus on addressing the following three essential ones: (1) managing diverse conversational purposes, (2) facilitating effective memory management for long-term conversations, and (3) automating the evaluation of machine-generated dialog responses. To overcome these challenges, we introduce HOKIEBOT, an open-domain chatbot developed for the Alexa Prize SocialBot Grand Challenge 5. HOKIEBOT employs a diverse set of dialog responders, including retrieval-based, neural network-based, and large language model-based (such as BlenderBot and Alpaca) to cater to a broad spectrum of user requests. To enhance engagement and foster long-term conversations, we introduce a novel topic-aware responder that keeps track of user preferences toward various topics from previous interactions, stores them in memory, and dynamically utilizes them to generate consistent responses. Additionally, we investigate various ranking strategies to evaluate and select the most suitable responses from a diverse array of candidates. The integration of these components enables HOKIEBOT to produce user-preferred responses and maintain consistency in long-term interactions, thereby offering an improved conversational experience across a wide range of conversational topics.

1 Introduction

As virtual assistants and smart speakers gain popularity and become more accessible, dialogue systems have assumed a pivotal role in the realm of natural language processing. Recent advancements in state-of-the-art dialogue systems, such as BlenderBot 3 [Shuster et al., 2022], ChatGPT [OpenAI, 2023], and Vicuna [Chiang et al., 2023], have demonstrated remarkable progress in generating fluent and coherent responses. These dialog systems are designed with the primary objective of creating agents capable of engaging in human-like conversations across a wide range of scenarios, producing consistent and user-preferred responses for both short-term and long-term dialogues.

In this work, we introduce **HOKIEBOT**, an innovative open-domain chatbot developed for the Alexa Prize SocialBot Grand Challenge 5 Johnston et al. [2023]. Our chatbot is specifically designed to address the following three crucial challenges: (1) the capacity of managing diverse conversational purposes effectively, (2) efficient memory management to facilitate long-term conversations, and (3) automating the evaluation of machine-generated dialogue responses.

5th Proceedings of Alexa Prize SocialBot Grand Challenge (Alexa Prize 2023).

Diverse Conversational Purposes Open-domain conversational agents or chatbots, such as Amazon Alexa [Goel et al., 2019, Krause et al., 2017, Gopalakrishnan et al., 2019], are designed and expected to seamlessly blend entertaining wit and knowledge while making the users feel heard and engaged. However, catering to the broad range of conversational topics and human requests, such as *general chat*, *seeking information*, and *in-depth discussions of events or entities*, presents a significant challenge in training a single conversational agent to fulfill all these varied purposes [Roller et al., 2021]. To support diverse conversational purposes, our system employs a combination of retrieval-based and neural-based chatbot responders, as discussed in Section 3, to effectively cover various conversational objectives.

Memory Management for Long-term Conversations Preserving an effective memory for long-term conversations is essential for generating consistent responses across multiple interactions and adapting to users' preferences. While recent studies have made efforts in personalizing dialogue systems for long-term conversations [Xu et al., 2022a,b, Bae et al., 2022], they primarily focus on how to efficiently store memory obtained from previous conversations, with less attention to guiding the dialogue model in using the memory to generate consistent and user-preferred responses, and dynamically updating the long-term memory.

In this work, we propose PERSONADIAL (Section 3.2), a novel framework that dynamically stores users' preferences and personalities, leveraging this information to guide the dialogue model. This approach ensures that the dialogue model generates responses that are consistent and aligned with users' interests, fostering a more engaging and long-term interaction with users.

Automatic Dialog Response Evaluation Evaluating the quality of machine-generated dialogue responses is a critical aspect in developing open-domain, general-purpose chatbots [Walker et al., 1997, Kamm, 1995]. Although human evaluation is typically considered the gold standard in evaluation, it requires significant labor effort and is not feasible for real-time evaluation of candidate responses generated by various models.

To address this challenge and enable real-time evaluation of machine-generated responses, we investigate various existing ranking strategies aimed at evaluating and selecting the most suitable responses from a vast array of potential candidates. Additionally, inspired by the recent success of instruction tuning that improves zero-shot performance on unseen tasks, we introduce INSTRUCTEVAL (Section 4.3), a novel unified automatic evaluation framework that is capable of following human instructions and evaluating responses on unseen customized aspects.

Our HOKIEBOT system is designed to be flexible, accommodating a wide range of conversational purposes, and makes use of effective memory management to produce consistent and user-preferred responses over numerous interactions. Additionally, our system incorporates an innovative automatic evaluation system for assessing and selecting the most suitable dialogue responses produced by the system.

2 System Overview

HOKIEBOT is an open-domain dialogue system developed utilizing the CoBot framework [Khatri et al., 2018]. The system is composed of multiple modules, each tailored to perform a specific task within the pipeline, as illustrated in Figure 1.

During each interaction, the user's input utterance undergoes a series of processing steps. First, the Global Intent Handler module manages the initial launch and termination intents, responding with an appropriate welcome or goodbye message based on the global user's intent. If the user intends to continue the conversation, the user utterance is then passed through the Automated Speech Recognition (ASR) Processor module, which converts the user's spoken words into written text. The ASR Processor also incorporates an offensive Content Filter that identifies any inappropriate content in the user utterance. If anything offensive is detected, the system triggers a topic redirection prompt to steer the conversation in a more suitable direction.

[2] Note that all exemplar conversations used in this report, including the utterances, characters, and names, are purely fictitious, i.e., they are first generated by large language models and then edited by humans. There is no correlation between the entities and real-world individuals.

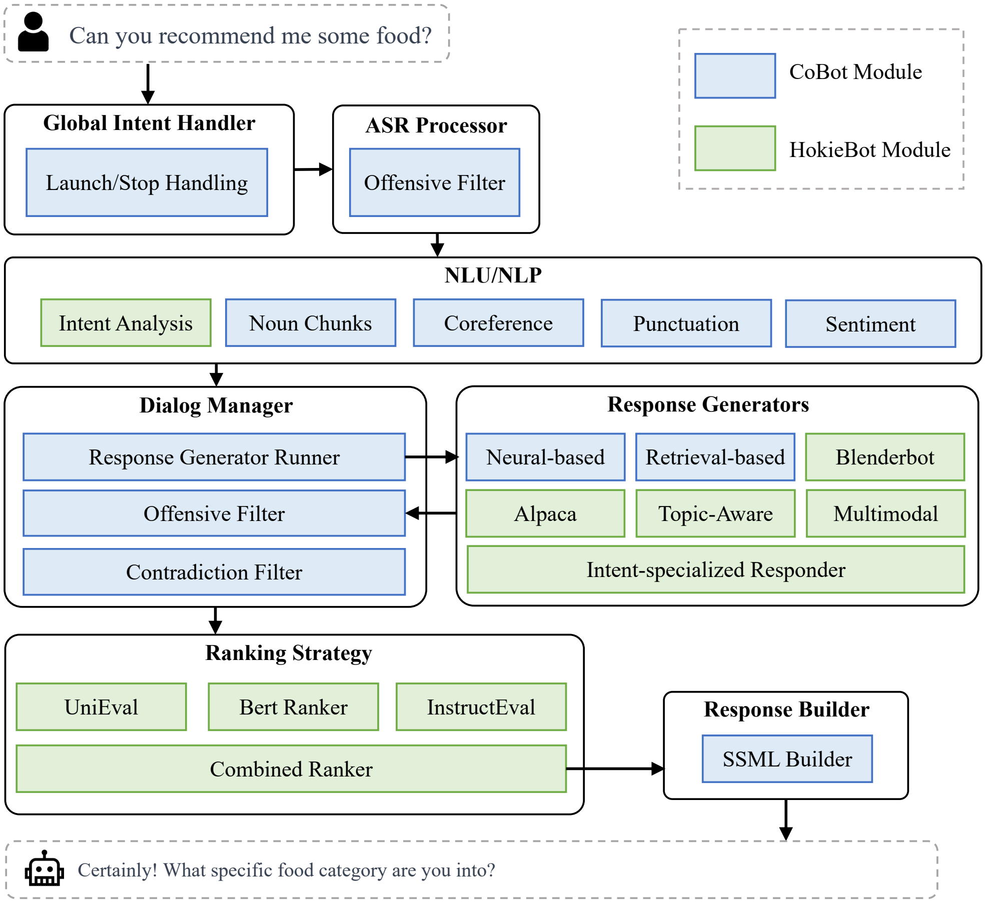


Figure 1: **System Overview of HokieBot.** The HokieBot is built on top of the CoBot framework. The blue boxes represent the default modules included in the CoBot framework, while the green boxes denote the modules unique to HokieBot.[2]

Following the ASR process, the NLU/NLP pipeline utilizes a collection of feature extractors to analyze the transcribed utterance in parallel. These extractors cover Intent, Noun Chunks, Coreference, Punctuation, and Sentiment detection, efficiently extracting valuable NLP features from the user's utterance. The identified features are then used in formulating appropriate responses.

The Dialog Manager module consists of a Response Generator Runner and a series of response filters. The Response Generator Runner launches multiple different responses (see Section 3 for details), creating a diverse set of candidate responses. Subsequently, these candidate responses undergo a series of filters designed to eliminate offensive, contradictory, or repetitive responses. The Ranking strategy takes the filtered responses from the Dialog Manager and employs a collection of ranking modules (see Section 4 for details) to assign scores to the candidate responses. These scores are then aggregated by a Combined Ranker to evaluate the overall score of each candidate response. Finally, the response with the highest ranking is selected and forwarded to the Response Builder, which prepares and formats the response for delivery to the user.

3 Response Generators

3.1 Conversational Responders

As we advance towards an era of large-scale general models capable of handling a wide range of tasks, it's important to recognize that specialized or domain-specific models still hold an advantage in terms of both specialty and computation efficiency. In particular, for real-time conversations, smaller and more diverse models are preferred due to their ability to deliver specialized responses tailored to specific situations. For instance, when a user seeks knowledge-based information, a knowledge responder would better suit the user's requirements, while in other scenarios, an empathy responder would be more appropriate for providing comfort and understanding to the user.

Open-domain conversational agents such as Amazon Alexa [Goel et al., 2019, Krause et al., 2017, Gopalakrishnan et al., 2019] are designed to blend entertaining wit and knowledge seamlessly while ensuring engagement and responsiveness. However, the broad range of conversational topics and human requests, such as *general chat*, *seeking information*, *discussing events or entities in depth*, and more, pose significant challenges in training a single conversational agent to cater to all purposes [Roller et al., 2021, Liu et al., 2023a]. One common approach employed by existing open-domain conversational agents involves employing a variety of responder modules, each trained with a specific objective, such as *chit-chat* [Chiu et al., 2022, Sun et al., 2021], *question-answering*, or *knowledge-grounded conversation generation* [Sun et al., 2022b]. The system then selects the optimal response from various candidates generated by these responders using a ranking strategy. HokieBot follows similar ideas and integrates a variety of expert responders, including both neural-based [Zhang et al., 2020, Shuster et al., 2022, Xu et al., 2023a] and retrieval-based [Henderson et al., 2017, Wu et al., 2019, Pan et al., 2021, Jhan et al., 2021] chatbot responders.

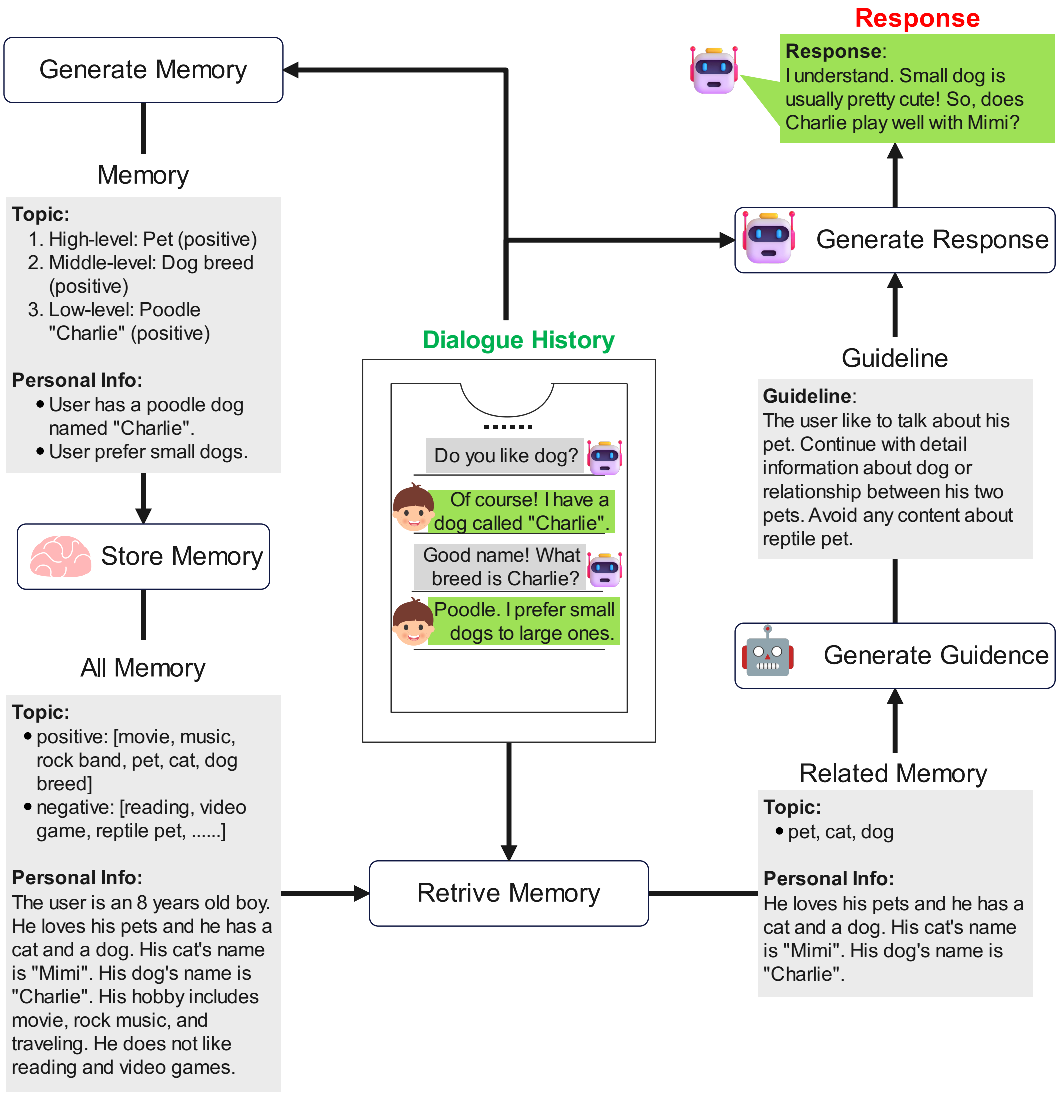
Retrieval-based Responder HokieBot fully leverages the following retrieval-based responders supported by the Cobot framework [Khatri et al., 2018]: **Fun Fact responder** is a responder using the knowledge retrieval API to return a fun fact about an entertainer. **News responder** uses the News Retrieval API to retrieve news about the user utterance. **Greeting responder** randomly selects a greeting message from a list of greeting templates.

Neural-based Responder We also employ a series of neural-based responders to generate candidate responses to the user utterance. The first two are **Topic NRG** and a **QA responder**, which are both conversational modules supported by the Cobot system. Topic NRG is trained by multi-task learning with two objectives for dialog topic and intent classification tasks. It classifies a dialog into 10 predefined topic categories, such as *Entertainment*, *Sports*, *Politics*, or *Other*. The QA responder uses the Evi QA service to retrieve an answer to the input question and outputs the answer if there is a high enough confidence score indicating that the answer is relevant. Additionally, we also develop several large language model-based responders, including **Blenderbot-400M-distill**, **Blenderbot-1B-distill**, **AlexaTM** [Soltan et al., 2022], and **Alpaca-LoRA-7B** which is implemented with fp16 to best meet the latency requirement. Furthermore, we incorporate a **multimodal dialogue responder** to enable the generation of multimodal responses that contain both visual and text content.

3.2 Topic Tracker and Topic-aware Responder

Motivation With the increasing popularity and accessibility of virtual assistants and smart speakers, dialogue systems have emerged as important conversation partners in everyday life. The ability to generate consistent responses across numerous interactions and adapt to users' preferences is crucial for sustaining long-term engagement and building long-term friendships. State-of-the-art dialogue systems such as BlenderBot 3 [Shuster et al., 2022], ChatGPT [OpenAI, 2023], and Vicuna [Chiang et al., 2023] have made a remarkable improvement on generating fluent and coherent responses. However, due to their fixed input token length, they are unable to consider the entire conversation history across multiple sections and hence mainly focus on single-section conversations.

Recent efforts have been made to personalize dialogue systems for long-term conversations. Xu et al. [2022a] suggests summarizing conversation sections to effectively store them and Xu et al. [2022b] proposes to only store responses that embody both the user's and chatbot's persona to improve long-term persona dialogue. However, the accumulation of stored memory and persona responses can lead to an infinite growth problem. To overcome this issue, Bae et al. [2022] developed a dialogue long-term memory management technique that selectively removes invalidated or redundant user information from memory. While the above methods primarily focus on efficiently storing persona information obtained from previous conversations into memory, they overlook two other more important questions: (1) *how to effectively guide the dialogue model to generate consistent and user-preferred responses by utilizing the memory?* and (2) *how to better construct and dynamically update the long-term memory?*



Response:
I understand. Small dog is usually pretty cute! So, does Charlie play well with Mimi?

Topic:
1. High-level: Pet (positive)
2. Middle-level: Dog breed (positive)
3. Low-level: Poodle "Charlie" (positive)

Personal Info:
- User has a poodle dog named "Charlie".
- User prefer small dogs.

......

Do you like dog?

Of course! I have a dog called "Charlie".

Good name! What breed is Charlie?

Poodle. I prefer small dogs to large ones.

Guideline:
The user like to talk about his pet. Continue with detail information about dog or relationship between his two pets. Avoid any content about reptile pet.

Topic:
- positive: [movie, music, rock band, pet, cat, dog breed]
- negative: [reading, video game, reptile pet, ...]

Personal Info:
The user is an 8 years old boy. He loves his pets and he has a cat and a dog. His cat's name is "Mimi". His dog's name is "Charlie". His hobby includes movie, rock music, and traveling. He does not like reading and video games.

Topic:
- pet, cat, dog

Personal Info:
He loves his pets and he has a cat and a dog. His cat's name is "Mimi". His dog's name is "Charlie".

Figure 2: **The overview of the PersonaDial framework.**

Approach Overview To address the aforementioned problems, we propose PersonaDial, a novel framework for automatically generating personalized responses based on previous interactions. Specifically, PersonaDial consists of five components: (1) a *memory module* that stores user preferences, use cases, and identity of the dialogue system; (2) a *memory generation module* that takes in conversation history and incorporates new user information into the memory module; (3) a *memory retrieval module* that takes dialogue history as input and retrieves potentially useful knowledge from the memory for response generation; (4) a *guidance generation module* that utilizes both dialogue history and retrieved user-specific knowledge to produce a guideline for generating next response; and (5) a *dialogue module* that simultaneously considers the dialogue history and the guideline to generate a new response. We show the overall architecture with a concrete example in Figure 2.

Open-domain Conversation Data Generation We focus on investigating the impact of user preference toward various topics on dialogue generation. To support this research, we collect a novel open-domain dialogue dataset comprising around 7,000 conversations across 44 topics. All conversations are automatically generated by ChatGPT, and the topics with user preferences for each utterance are also automatically labeled by ChatGPT by leveraging its in-context learning capability.

Specifically, as shown in Figure 3, we prompt ChatGPT with five examples of annotated dialogues while each example consists of four components:

1. A **human-like conversation** containing up to 10 conversational turns. Considering the input length constraints of Large Language Models (LLMs), we imposed a limit of 10 conversational turns per dialogue sample.
2. **Topic annotations** of the last conversational turn, dissected at three levels of granularity. We categorize the topic of each conversation turn into three levels: **high**, **middle**, and **low**. **High-level** topics correspond to expansive categories that encapsulate a variety of subtopics and experiences (e.g., *sports*, *music*, *art*). **Middle-level** topics relate to more specific subtopics within the high-level categories, possessing a heightened focus and specificity (e.g., *football*, *rock bands*, *painting*). **Low-level topics** pertain to specific entities, such as people, places, or things (e.g., *Cristiano Ronaldo*, *The Rolling Stones*, *Vincent van Gogh*).
3. **Binary preference annotations** for the user with respect to each topic. We ask ChatGPT to provide the user preference for each topic, contingent on the conversational context, signaling whether the user prefers to continue with the current topic or divert to an alternative.
4. (optional) a succinct **summary of preceding dialogue topics** and a **suggested topic** for the subsequent response. Predicated on the preference of preceding topics, we further instruct ChatGPT to summarize the user's intent and provide a topic suggestion for subsequent dialogue.

Although the utterances in the prompt were not specific quotes from users, their style and depth were determined based on observations of common types of user utterances while following common topic trends (movies, food, books, and music). Figure 4 shows several data examples from our generated open-domain conversation dataset.

Memory Generation (MG) We fine-tune a dialogue model such that given the conversation history as input, the model can generate the current topics at three levels and their human preferences. Note that the human preference for each topic is independent. It is common for individuals to be interested in a high-level topic but not a specific middle or low-level topic. Figure 5 shows some examples of topic extraction at three levels. We define three sets of topics in the memory: `Rejected`, `Interested`, and `Potential Interested`. `Rejected` refers to the topics that human is not interested in. `Interested` topics are the ones for which users chat with our bot normally. If there is no negative sentiment detected from the user, we regard the topic as an `Interested` one. `Potential Interested` topics are the ones inferred based on `Interested` topics. When users are interested in a lower-level topic, we use the LLMs to predict several `Potential Interested` topics with the same higher-level topic, which are later used as candidates for shifting topics. We categorize the topics and preferences into these three sets. However, human preferences even on the same topic may still change. We will keep updating the memory with the latest human preference.

Guidance Generation (GG) Previous work [Gupta et al., 2022b] has shown that dialogue models can effectively follow a set of pre-defined guidelines to generate corresponding responses. Due to the diversity of topics, instead of using a fixed set of guidelines, we propose a guideline generation module that simultaneously considers the retrieved topics and dialogue history to generate guidelines aligned with user preference. Specifically, the guideline generation module will first decide if the chatbot should stick to the current topic or not based on the human preference. If so, it will invoke an LLM by taking the n-turns of conversation history, their topics and human preference as input to generate a guideline to guide the responder to follow the previous topic; if not, the module will first retrieve some topics candidate from `Potential Interested` set, then invoke the LLM with these candidates as input to generate a guideline and encourage the chatbot to shift to a new topic.

Dialogue Responder (DR) Our Dialogue Responder is developed by leveraging instruction tuning[Gupta et al., 2022b, Xu et al., 2023b, Qi et al., 2023] where the output from the Guidance Generation module is viewed as the instruction to response generation. Specifically, the Dialogue Responder takes in the output from the **Guidance Generation** module as well as the ongoing conversation history and generates an appropriate response by following the guideline.

Implementation Details We implement the topic extractor based on InstructDial [Gupta et al., 2022a] and only store the topic as memory to help generate more robust responses. To extract

Here are 5 examples of a conversation containing 3 pieces: the conversation history, user topic preferences, and the guidelines for a chat assistant. Each of these are separated by the "|" token.

1. B:Do you like eating food? A:I love eating most kinds of food. B:What is something that you do not like? A:I do not like mexican food. |{"high-level": {"topic": "food", "if_interest": "yes"}, "middle-level": {"topic": "Mexican food", "if_interest": "no"}} | The user is interested in talking about food. They do not like Mexican food, so talk about another type of food.
2. B:What do you like to do? A:I like listening to rock n roll music. I really like The Beatles and Elvis Presely. | {"high-level": {"topic": "music", "if_interest": "yes"}, "middle-level": {"topic": "rock n roll", "if_interest": "yes"}} {"high-level": {"topic": "music", "if_interest": "yes"}, "middle-level": {"topic": "bands/artists", "if_interest": "yes"}, "low-level": {"topic": "The Beatles/Elvis Presley", "if_interest": "yes"}} | The user likes to listen to music. They like the rock n roll genre. They like the band 'The Beatles' and the artist 'Elvis Presely'. Tell them about other rock n roll artists similar to 'The Beatles' and 'Elvis Presely'.
3. B:What is a hobby that you like? A:I like reading. I like reading fantasy books, but I do not like 'Dune'. | {"high-level": {"topic": "reading", "if_interest": "yes"}} {"high-level": {"topic": "reading", "if_interest": "yes"}, "middle-level": {"topic": "genre", "if_interest": "yes"}, "low-level": {"topic": "fantasy", "if_interest": "yes"}} {"high-level": {"topic": "reading", "if_interest": "yes"}, "middle-level": {"topic": "book", "if_interest": "unknow"}, "low-level": {"topic": "Dune", "if_interest": "no"}} | The user likes to reed books. They specifically like to read fantasy books. They are not interested in reading the book 'Dune'. Talk to them about any other potential books that they like reading.
4. A:I do not like sushi. B:What kind of food do you like? A:I like Italian and Mexican cuisine. B:What Italian and Mexican dishes are your favorite? A:Lasagna, spaghetti bolognese, tacos, and burritos. | {"high-level": {"topic": "food", "if_interest": "no"}, "low-level": {"topic": "sushi", "if_interest": "no"}} {"high-level": {"topic": "food", "if_interest": "yes"}, "middle-level": {"topic": "cuisine", "if_interest": "yes"}, "low-level": {"topic": ["Italian", "Mexican"], "if_interest": "yes"}} {"high-level": {"topic": "food", "if_interest": "yes"}, "middle-level": {"topic": "Italian cuisine", "if_interest": "yes"}, "low-level": {"topic": ["lasagna", "spaghetti bolognese"], "if_interest": "yes"}} {"high-level": {"topic": "food", "if_interest": "yes"}, "middle-level": {"topic": "Mexican cuisine", "if_interest": "yes"}, "low-level": {"topic": ["tacos", "burritos"], "if_interest": "yes"}} | The user does not like the food sushi. However, they like Italian and Mexican cuisine. They specifically like lasagna, spaghetti bolognese, tacos, and burritos. Ask them about some other Italian or Mexican cuisine dishes that they like or that you think they would like to try.
5. A:TV series are not my favorite, but I do like comedy. B:Do you like Game of Thrones? A: No. B:What comedies do you like? A:I like the office. My favorite moment from it is the dinner party episode. B:What is another comedy that you like? A:I also really enjoy Friends. | {"high-level": {"topic": "TV series", "if_interest": "yes"}, "middle-level": {"topic": "Game of Thrones", "if_interest": "no"}} {"high-level": {"topic": "TV series", "if_interest": "no"}} {"high-level": {"topic": "TV series", "if_interest": "yes"}, "middle-level": {"topic": "comedies", "if_interest": "yes"}, "low-level": {"topic": "The Office", "if_interest": "yes"}} {"high-level": {"topic": "TV series", "if_interest": "yes"}, "middle-level": {"topic": "favorite moment", "if_interest": "yes"}, "low-level": {"topic": "Dinner Party episode", "if_interest": "yes"}} {"high-level": {"topic": "TV series", "if_interest": "yes"}, "middle-level": {"topic": "comedies", "if_interest": "yes"}, "low-level": {"topic": "Friends", "if_interest": "yes"}} | The user is generally not interested in TV series. They specifically do not like 'Game of Thrones'. They are however, interested in the comedy 'The Office'. Their favorite moment from the series is the 'Dinner Party' episode. Another TV series they like is 'Friends'. Ask the user why they don't like most TV series other than comedies.

Generate 20 new and unique examples similar to the provided 5. Include all 3 pieces: the conversation history, the topic preferences, and the guidelines for the conversation. Make each generated example different from each other, but make sure to follow the format seen in the previous 5 examples. Make sure that the examples alternate how many preferences are present in each generation.

Figure 3: Example of an in-context learning prompt for generating the open-domain conversational data.

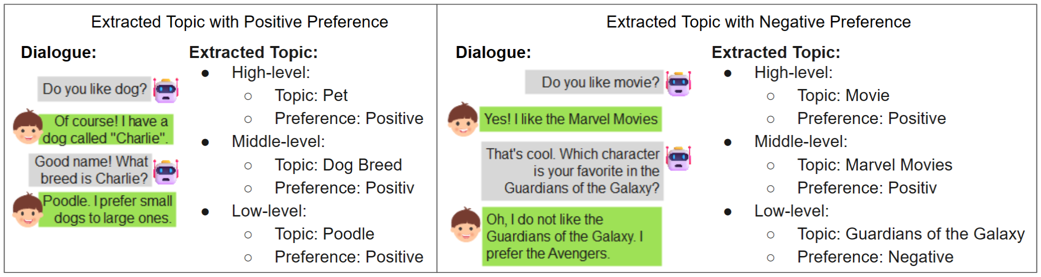
topic and human preference in our pre-defined format, the `Dial-BART0` checkpoint is fine-tuned on our generated open-domain conversational dataset, spanning 44 high-level topics. For guidance generation, we utilize Flan-T5 [Chung et al., 2022] which is fine-tuned on a subset of our generated dataset involving 981 pairs of conversations and guidelines. For the response generation module, we fine-tune the `blenderbot-400M-distill` checkpoint using a subset of our generated dataset comprising 1959 triples, each consisting of a conversation, a guideline, and a response. All training processes use the AdamW optimizer with a learning rate of 1e-6, running over 20 epochs.

Example 1:
Conversation:
...... A: Do you like sports? B: Yes, I do. I particularly enjoy basketball.
Topics: {“high-level”:{“topic”:“sports”,“if_interest”:“yes”},“middle-level {“topic”: “basketball”, “if_interest”: “yes”}}
Guidance: The user likes sports and basketball. Talk to them about their favorite basketball teams or players.

Example 2:
Conversation:
...... A:I like playing video games. B:What type of video games do you enjoy? A:I like playing RPGs and action/adventure games.
Topics: {“high-level”: {“topic”: “video games”, “if_interest”: “yes”}, “middle-level”: {“topic”: “genre”, “if_interest”: “yes”}, “low-level”: {“topic”: [“RPG”, “action/adventure”], “if_interest”: “yes”}}
Guidance: The user enjoys playing video games in the RPG and action/adventure genres. Ask about their favorite game or suggest a new one they may enjoy.

Example 3:
Conversation:
...... A: I’m not interested in politics. B: What other current events are you interested in? A: I enjoy following the stock market.
Topics: {“high-level”: {“topic”: “current events”, “if_interest”: “yes”}, “middle-level”: {“topic”: “politics”, “if_interest”: “no”}} {“high-level”: {“topic”: “current events”, “if_interest”: “yes”}, “middle-level”: {“topic”: “finance”, “if_interest”: “yes”}, “low-level”: {“topic”: “stock market”, “if_interest”: “yes”}}
Guidance: The user is not interested in politics, but they like following the stock market. Ask them about their knowledge of finance and suggest similar topics they might want to know about.

Figure 4: Examples of generated open-domain conversation data, each with annotations of topics, user preferences, and suggested new topics.



Extracted Topic with Positive Preference

Dialogue:

Do you like dog?

Of course! I have a dog called "Charlie".

Good name! What breed is Charlie?

Poodle. I prefer small dogs to large ones.

Extracted Topic:

- High-level:
  - Topic: Pet
  - Preference: Positive
- Middle-level:
  - Topic: Dog Breed
  - Preference: Positiv
- Low-level:
  - Topic: Poodle
  - Preference: Positive

Extracted Topic with Negative Preference

Dialogue:

Do you like movie?

Yes! I like the Marvel Movies

That's cool. Which character is your favorite in the Guardians of the Galaxy?

Oh, I do not like the Guardians of the Galaxy. I prefer the Avengers.

Extracted Topic:

- High-level:
  - Topic: Movie
  - Preference: Positive
- Middle-level:
  - Topic: Marvel Movies
  - Preference: Positiv
- Low-level:
  - Topic: Guardians of the Galaxy
  - Preference: Negative

Figure 5: **Examples of three level topics in PersonaDial**

3.3 Multimodal Dialogue Responder

Motivation and Background The significance of *Multimodal Dialogue Response Generation* lies in enabling an exceptional intelligent conversational agent to transcend plain text communication and embrace the power of perceiving and conveying the real visual physical world. In human conversations, images effortlessly convey profound visual perceptions that are challenging to express solely through plain text. Considering this, recent work [Sun et al., 2022a, Firdaus et al., 2021] begin to explore the *Multimodal Dialogue Response Generation* either by generation-based methods [Sun et al., 2022a] or retrieval-based methods [Koh et al., 2023]. Several multimodal dialogue datasets [Feng et al., 2022, Firdaus et al., 2021] are created, which can support this line of research. In this work, we treat it as a multimodal entity linking task [Yao et al., 2023, Wang et al., 2023, 2022] to ground and link the mention within the generated response to an entity in a target knowledge base (KB) and combine the image of the corresponding entity with the textual response as our multimodal response.

Problem Formulation We formulate the entity linking-based multimodal response generation task as follows. Given a generated textual response t_r containing an entity mention m_r, e.g., “*Image Dragons*”, we aim to link the mention to a unique entity in the target KB, e.g., “*Rock band - Image Dragons*”, and combine the entity image with the textual response to be the multimodal response.

Each entity e_j in the KB is described with a textual description d_{e_j}, a name t_{e_j}, and several images $\bar{\mathbf{V}}_{e_j} = \{v^0_{e_j}, ..., v^h_{e_j}\}$.

Approach Details Figure. 6 shows the architecture for our multimodal responder. Given a response generated by HOKIEBOT, we first retrieve K (K=10) candidate entities from the KB whose lemmatization-based root form of n-gram span ($n \in \{1, 2, 3, 4\}$) matches any named entities in the HOKIEBOT's response. We then apply CLIP [Radford et al., 2021] as the encoder to obtain an overall representation of HOKIEBOT's response and a representation for each entity image, and further compute the cosine similarity between these representations. We sort all the images from the K candidate entities based on the cosine similarity scores and take the top-1 image as the aligned image to form the multimodal response. We use Wikidiverse [Wang et al., 2022], which contains 16M entities, as our target knowledge base.

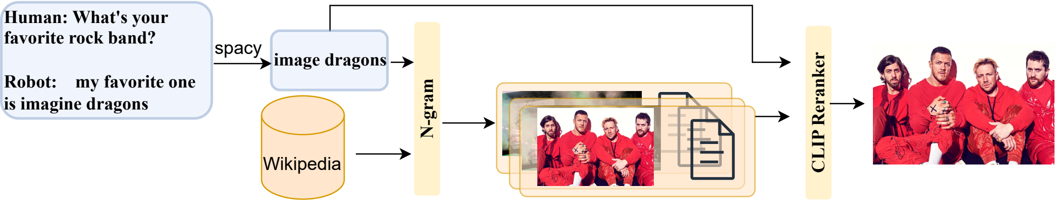

Human: What's your favorite rock band?
Robot: my favorite one is imagine dragons
spacy
image dragons
Wikipedia
N-gram
CLIP Reranker

Figure 6: The architecture of multimodal responder. The mention within the text is linked to one entity within the knowledge base.

3.4 Intent Specialized Responder

Inappropriate User Request The user may request the bot to execute inappropriate operations. For example, the user may ask the bot to *play music* while the current Socialbot environment has no access to other applications like the music player. On the other hand, the user may ask the bot to *turn on the light* while the bot can not execute physical actions so far. To handle this issue, we design several regular expressions to detect these inappropriate user requests and reply with a template to explain that our bot can not help with these actions but we are happy to discuss other topics.

Stop Intent Detection When users hope to stop the conversation, they can say “stop” to exit it. However, the user may mistakenly use other commands like “cancel” when they want to stop the conversation. To address the user's stop intent, we apply a regular expression to detect the sentences like “Alexa, cancel” and remind the user that they can say “stop” to end the conversation.

4 Ranking Strategy

As our system consists of a variety of response generators, a response ranking strategy is crucial for selecting the most suitable response. In this section, we explore various existing dialog response evaluators and further propose a more robust, generalizable, and innovative evaluator. Finally, we employ a combined ranker to integrate the ranking scores from all the evaluators for selecting the optimal response while considering the dialogue history.

4.1 Bert Ranker

Bert Ranker is provided as part of Cobot Toolkit Service, aiming to select the most appropriate response from a set of responders. It was finetuned using a BERT-base [Devlin et al., 2019] model on a dataset consisting of multi-turn dialogs from Alexa Prize logs, where each turn has a list of relevant and irrelevant responses. The model is trained as a binary classifier to score each response as relevant or irrelevant and brings a 5% absolute increase in Recall@1. One limitation of the Bert ranker is that it only outputs an overall score for each candidate response without providing fine-grained evaluation over different dialogue aspects, such as *coherence*, *groundedness* and so on.

4.2 UniEval Ranker

UniEval [Zhong et al., 2022] is an advanced and comprehensive unified multi-dimensional evaluator, designed to assess diverse facets of text generation. This evaluation framework formulates the evaluation process as a Boolean Question Answering (QA) task. UniEval applies continual learning [Parisi et al., 2019], a learning technique that is widely applied in various NLP tasks [Madotto et al., 2021, Liu et al., 2022, Liu and Huang, 2023], to sequentially learn various evaluation aspects. It is worth noting that UniEval provides distinctive evaluators fine-tuned specifically for different types of generation tasks, encompassing areas such as dialogue generation, summarization, and beyond. For the purposes of this work, we employ the UniEval dialogue evaluator to assess aspects of *naturalness*, *coherence*, *engagingness*, *groundedness*, and *understandability*.

Repetitiveness Evaluation We further add the aspect of "*repetitiveness*" into the UniEval module since we have noticed a certain amount of repetitiveness within the responses from the Topic NRG and other responders. It includes *Inter-turn Repetitiveness*, where the bot may ask the same question which has been asked in the previous dialogue turns, and *Intra-turn Repetiviveness*, where the bot repeats the same sentence multiple times in response. To solve these two issues, we incorporate a repetitiveness checking for each generated response by splitting the generated response R and dialogue history D into sentences and assign a low score to the corresponding response if there is the same sentence within the response, *Intra-turn Repetiviveness*, or there is the same sentence between response R and dialogue history D. The response with a low score will be excluded from the final response.

4.3 InstructEval

To develop an evaluator that can perform the evaluation on unseen customized aspects with high correlation with humans, we develop a scalable and customizable automatic evaluation approach that can be universally and flexibly applied to various aspects (e.g., *naturalness* and *engagingness*). Inspired by the recent success of instruction tuning that it can remarkably improve zero-shot performance on unseen tasks, we propose to finetune a pre-trained Large Language Model (LLM) on a wide range of tasks such as *scoring, comparison, and ranking* to explicitly align the model with human preferences. During the test time, the instruction-tuned model performs fine-grained evaluation given the aspect and its definition specified in the instruction even if the aspect is not present during training.

Instruction Tuning on Diverse Forms of Feedback Prior studies have shown that the number and diversity of instruction tuning tasks are critical aspects of the generalization ability of the model [Wei et al., 2022, Chung et al., 2022]. Motivated by this, we propose to construct a comprehensive and diverse collection of tasks that are specifically tailored for dialogue evaluation. In the following, we elaborate on how we construct the tasks used for instruction tuning and how they can benefit our evaluator. **(1) Scoring:** the first type of task is training the model to follow scoring criteria and directly score a response on a certain aspect specified in the criteria. For example, we construct an instruction *"A score of 0 indicates the response is not engaging. A score of 1 indicates the response is engaging. Select a score based on the engagingness of the response."*; **(2) Comparison:** the second type of task is to ask the model to select a better response from the two responses based on the context and the specified aspect. By learning from the comparison, the model can be aware of which response has better quality; **(3) Ranking:** to enable the model to learn from more fine-grained and subtle differences in human preference, we further extend the comparison task into a ranking task, where the model needs to predict a ranking of a set of responses, and; **(4) Question answering:** we also formulate the dialogue evaluation as a Boolean question answering task following Zhong et al. [2022]. For example, we asked *"Does this response sound natural?"* and let the model predict *"Yes"* or *"No"*. We adopt this type of task to evaluate the dialogue during the test time. To enhance the diversity of instruction tuning tasks, we employ evaluations from a variety of text generation tasks, including dialogue generation, summarization, and table-to-text generation. We train the evaluator on various aspects (e.g., *relevance, coherence, fluency,* and *diversity*) and perform a meta-evaluation in terms of the correlation with human evaluation on both seen and unseen aspects (e.g., *informativeness* and *groundedness*). We present our proposed framework in Figure 7.

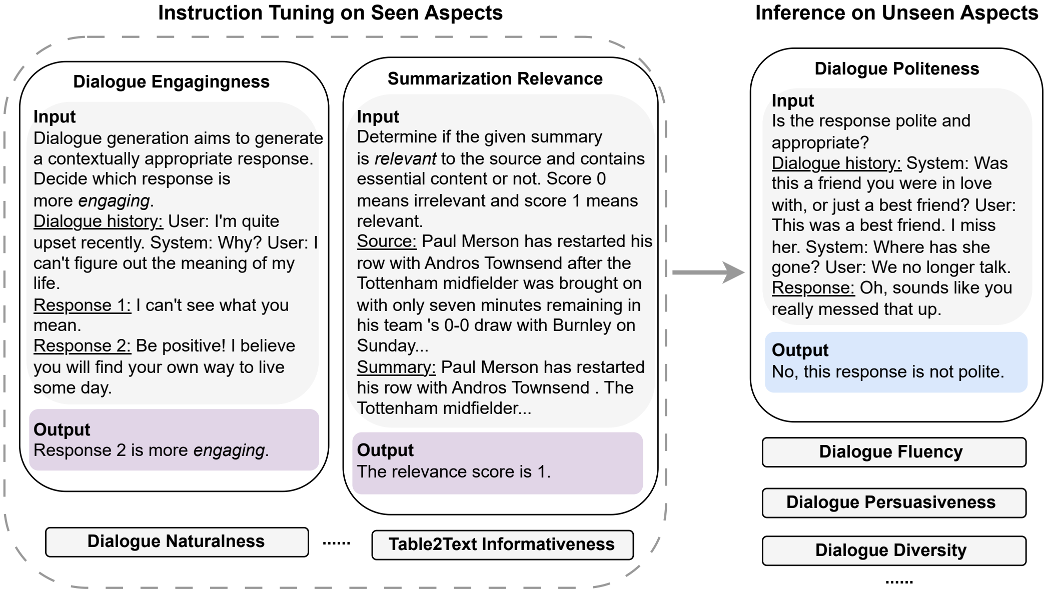

Instruction Tuning on Seen Aspects
Inference on Unseen Aspects
Dialogue Engagingness
Input
Dialogue generation aims to generate a contextually appropriate response. Decide which response is more *engaging*.
Dialogue history: User: I'm quite upset recently. System: Why? User: I can't figure out the meaning of my life.
Response 1: I can't see what you mean.
Response 2: Be positive! I believe you will find your own way to live some day.
Output
Response 2 is more *engaging*.
Summarization Relevance
Input
Determine if the given summary is *relevant* to the source and contains essential content or not. Score 0 means irrelevant and score 1 means relevant.
Source: Paul Merson has restarted his row with Andros Townsend after the Tottenham midfielder was brought on with only seven minutes remaining in his team 's 0-0 draw with Burnley on Sunday...
Summary: Paul Merson has restarted his row with Andros Townsend . The Tottenham midfielder...
Output
The relevance score is 1.
Dialogue Naturalness
......
Table2Text Informativeness
Dialogue Politeness
Input
Is the response polite and appropriate?
Dialogue history: System: Was this a friend you were in love with, or just a best friend? User: This was a best friend. I miss her. System: Where has she gone? User: We no longer talk.
Response: Oh, sounds like you really messed that up.
Output
No, this response is not polite.
Dialogue Fluency
Dialogue Persuasiveness
Dialogue Diversity
......

Figure 7: **Illustration of our InstructEval framework.** We first instruction-finetune the LLM on a diverse range of tasks so that the model can learn to follow the instruction and perform the evaluation on specified aspects. During inference, we can directly apply the model to unseen evaluation aspects such as *politeness* and *diversity*.

4.4 Ranking Aggregation

We sum the scores from Bert Ranker, UniEval Ranker, and InstructEval Ranker to form the final ranking score, while the response with the highest final ranking score will be chosen as the most appropriate response. Our error analysis shows that Blenderbot and ATM20B often lead to better responses, while our ranker may not always assign the highest scores to these two modules. Considering it, we intentionally increase the combined score of Blenderbot by 0.15, increase the combined score of ATM20B by 0.1, decrease the combined score of ATM5B by 0.15, and decrease the combined score of the NRG model by 0.25.

5 Results and Analysis

5.1 Qualitative Analysis of Responder Outputs

To better evaluate the response quality of the proposed framework, we analyze low-score conversations (specifically conversations with scores 1 or 2) over six weeks[3] We summarize the main reasons for each conversion being not satisfying. They are essentially challenges in conversational systems. The distribution of the collected challenges is shown in Figure 8.

We summarize the causes for low-score conversations into the following 9 categories:

- **Short Conversations**. Based on the low-score conversation analysis, we found a significant amount of conversations that are scored very low while the conversations are extremely short, e.g., within 5 turns of conversations. Possible reasons include the user triggering the socialbot by mistake and thus immediately stopping the conversation, or a system crash due to unknown reasons.
- **Repeated Conversations**. The repeated conversation is the case that the bot repeated similar questions or conversations multiple times within one dialog. It happens because such a conversation will be assigned a higher coherence score as it aligns with the dialog history.

[3] We analyze three weeks of conversations in March, two weeks in April, and one week in June. All weeks are consecutive calendar days.

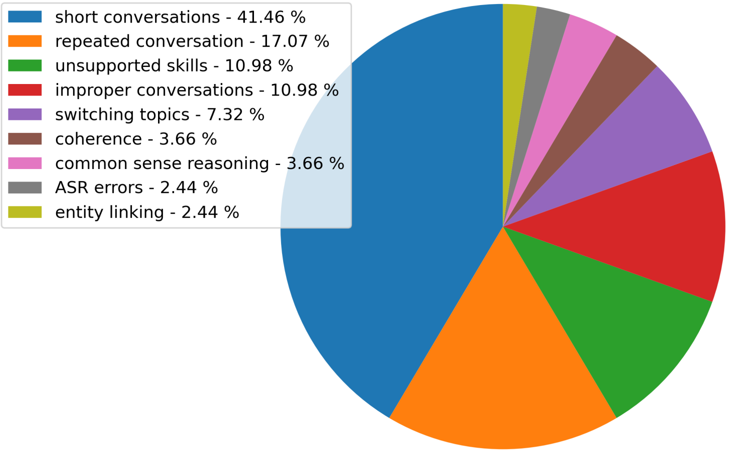


Figure 8: The distribution of the causes for low-score conversations.

- **Unsupported Skills**. Sometimes the user may ask for skills that are not yet supported by the system. For example, the user may ask the bot to sing a song or change voice. We expect such features to be supported in the future.
- **Improper Conversations**. Though the bot is able to filter out many improper inquiries such as sensitive information like a bank account and several polite response templates will be generated, there are still some cases where such filtering fails.
- **Switching Topics**. Another common issue happens when the user switches conversation topics. Responders are expected to follow the topic changes based on the conversation, e.g., from movies to favorite sports. Comparatively, BlenderBot performs better in the following topic changes, while others like TopicNRG and EmpthyNRG may stick to the previous conversation topics.
- **Coherence**. Coherence has been identified as an essential evaluation criterion in previous conversation evaluation metrics. Here, the coherence issue refers to the inconsistency in the bot identity. For example, the bot may claim to love tacos in one preceding response and later it may change its mind and claim to not be a fan of tacos.
- **Commonsense Reasoning**. Sometimes the bot cannot infer like a human. For example, the bot asks the user "*what is his/her favorite movie*", and the user replied "*The Green Mile*". Then the bot asked a follow-up question "*Have you seen it?*". This is clearly inappropriate from a human perspective because we know that if the user liked the movie, the user must have seen it.
- **ASR Errors**. ASR errors have been identified as another issue in the current system. For example, "*cat*" may be recognized as "*car*". Such speech recognition errors are vital because they can dramatically switch the conversation topic from animals to vehicles.
- **Entity linking Errors**. The system may fail to link the entity mentions to its actual meaning.

5.2 Efficacy of Topic Tracking

Our experiments demonstrate PERSONADIAL's proficiency in identifying conversation topics and generating personally preferred responses. Qualitative analysis that compares responses from PERSONADIAL and BlenderBot is provided in Figure 9.

In comparison to BlenderBot, PERSONADIAL demonstrates a more refined and gradual approach in transitioning dialogue topics, resulting in a more coherent and engaging conversation. When the user shows no apparent resistance to the current topic, PERSONADIAL tends to maintain focus on the current topic, fostering a more fine-grained discussion rather than abruptly veering off to unrelated topics. As illustrated in case 1 in Figure 9 where the conversation revolves around a purple monster, the baseline model, driven by its objective to promote diverse discussions, shifts the discussion toward pets, while PERSONADIAL steers the user to share more memories related to the color purple, thereby preserving continuity and natural flow.

Case 1:

Conversation History:
......
HokieBot: Purple is a great color. Why are they afraid of purple?
User: I am scared of purple because it reminds me of a scary monster.

PersonaDial

Topic Detected:
{"high-level": {"topic": "fear_of_color", "if_interest": "yes"}, "middle- level": {"color": "purple", "if_interest": "yes"}}
Guideline:
The user is interested in talking about colors, especially purple. Ask them what their memories of purple are.
Response:
What's your most impressive memory of purple?

BlenderBot

Response:
That makes sense. Do you have any pets? I have a dog and a cat.

Case 2:

Conversation History:
......
User: The weather was very nice today.

PersonaDial

Topic Detected:
{"high-level": {"topic": "weather", "if_interest": "yes"}}
Guideline:
The user likes sunny and warm days with a gentle breeze. They prefer wearing a hat and sunglasses in sunny days. Ask them what their favorite outdoor activities are during sunny days.
Response:
That's interesting. What are some of the activities that you like to do on a sunny day?

BlenderBot

Response:
Yes, it was. It was a nice change of pace from the cold weather we've been having.

Case 3:

Conversation History:
......
HokieBot: What's your favorite rom-com comedy movie that you've watched?
User: I don't like rom com movies.

PersonaDial

Topic Detected:
{"high-level": {"topic": "movie", "if_interest": "yes"}, "middle-level": {"topic": "comedy movie", "if_interest": "yes"}, "low-level": {"topic": "rom-com", "if_interest": "no"}}
Guideline:
The user is not interested in movies but they do like rom-com comedy movies. They specifically like 'The Big Bang Theory' and 'The Theory of Everything'. Ask them about other comedy movies that they might like to watch.
Response:
That's interesting. Do you have any other movies or movies that you would recommend them to watch?

BlenderBot

Response:
Have you ever seen The Notebook? It's a romantic comedy that came out in 2007.

Figure 9: Example Responses from BlenderBot and PERSONADIAL.

Moreover, PersonaDial effectively handles situations when the user's input takes the form of an imperative sentence or when the current topic naturally reaches its conclusion. In such cases, PersonaDial proactively introduces new topics by posing questions, ensuring the dialogue's continuity. When introducing new topics, the system tends to be a logical extension or a subsidiary of the current subject, maintaining coherence at a higher level. For example, as depicted in case 2 (Figure 9), when the user initiates a discussion about the weather, PersonaDial suggests the topic of "*activity*" as an extension of "*good weather*", facilitating smooth conversational transitions. In contrast, BlenderBot tends to merely agree with the user's input, offering limited engagement.

Furthermore, PersonaDial excels in recognizing the user's disinterest in the current topic and promptly redirects the conversation to avoid dissatisfaction. To cater to the user's interests, the system selects new topics that are relevant sub-topics under the broader topics of interest. For example, in case 3 (Figure 9), when the user expresses dislike for "*rom-com comedy*" under the category of "*movies*", PersonaDial identifies this negative sentiment and seeks to redirect the conversation. Leveraging the user's positive sentiment towards "movies," PersonaDial navigates towards other film genres falling under the "*movie*" category. On the other hand, BlenderBot persists with the current topic, focusing solely on the presence of the keyword "*rom-com comedy*".

In summary, PersonaDial outperforms BlenderBot in maintaining smooth and coherent dialogue by skillfully handling topic transitions based on user engagement, proactively introducing new topics, and swiftly responding to user preferences. These strengths contribute to a more natural and enjoyable conversational experience for users.

5.3 Evaluation of Ranking Strategies

Meta Evaluation Datasets We meta-evaluate our evaluator across two benchmarks for dialogue evaluation, i.e., Topical-Chat [Gopalakrishnan et al., 2019] and FED [Zhang et al., 2020]. We report the performance on four aspects of Topical-Chat: `naturalness` (NAT), `coherence` (COH), `engagingness` (ENG), and `groundedness` (GRD). We hold out all 18 aspects in FED in instruction tuning to evaluate the performance on unseen aspects.

Implementation Details and Baselines We employ `FLAN-T5-large` [Chung et al., 2022] as the backbone LLM for our InstructEval. During instruction tuning, we sample part of synthetic data used in [Zhong et al., 2022] to enlarge and diversify our dataset. For G-Eval-3.5 and G-Eval-4 baselines, we report the performance in [Liu et al., 2023b] where GPT-3.5 and GPT-4 are used as the backbone, respectively. We reimplement GPTScore using `FLAN-T5-large` as the backbone since we cannot access GPT-3's output logits.

Metrics	**NAT**	**COH**	**ENG**	**GRD**	**Avg**
ROUGE-L [Lin, 2004]	0.146	0.203	0.300	0.327	0.244
BLEU-4 [Papineni et al., 2002]	0.175	0.235	0.316	0.310	0.259
METEOR [Banerjee and Lavie, 2005]	0.191	0.302	0.439	0.391	0.331
BERTScore [Zhang* et al., 2020]	0.209	0.233	0.335	0.317	0.273
USR [Mehri and Eskenazi, 2020b]	0.325	0.377	0.465	0.447	0.403
UniEval [Zhong et al., 2022]	0.514	0.613	0.605	0.575	0.577
GPTScore [Fu et al., 2023]	0.327	0.269	0.161	0.195	0.238
G-Eval-3.5 [Liu et al., 2023b]	0.539	0.544	**0.691**	0.567	0.585
G-Eval-4 [Liu et al., 2023b]	**0.565**	0.605	0.631	0.551	0.588
InstructEval w/o IT	0.174	0.155	0.192	0.293	0.204
InstructEval (Ours)	0.450	**0.646**	0.572	**0.706**	**0.593**

Table 1: Seen-aspect meta-evaluation in terms of turn-level Spearman correlation on Topical-Chat. The best results are highlighted in bold. "IT" refers to Instruction Tuning on our collected data.

Results We show the meta-evaluation results on Topical-Chat in Table 1, where the aspects are seen during instruction tuning. Our InstructEval outperforms previous state-of-the-art methods on `coherence` and `groundedness` aspects by 3.3% and 13.1% Spearman correlation, respectively. We hypothesize the reason why our evaluator performs better on COH and GRD while performing worse on NAT and ENG is that, the instruction tuning on summarization evaluation makes our approach better assesses objective aspects but it negatively impacts the evaluation of subjective aspects. We

Metrics	Dial	Turn	Avg
BARTScore [Yuan et al., 2021]	0.058	0.128	0.093
FED [Mehri and Eskenazi, 2020a]	0.204	0.119	0.162
DynaEval [Zhang et al., 2021]	0.398	0.256	0.327
UniEval [Zhong et al., 2022]	-0.310	0.187	-0.06
GPTScore [Fu et al., 2023]	0.113	0.099	0.106
INSTRUCTEVAL w/o IT	0.196	0.116	0.156
INSTRUCTEVAL (Ours)	**0.408**	**0.365**	**0.387**

Table 2: Unseen-aspect meta-evaluation in terms of Spearman correlation on FED. The best results are highlighted in bold. "IT" refers to Instruction Tuning on our collected data.

leave the analysis of potential negative transfer for future work. We also validate our method on the unseen aspect of FED and report the average performance across dialogue-level and turn-level aspects in Table 2. Our approach outperforms the baselines by a large margin. We also find that UniEval performs worst on dialogue-level evaluation. One plausible reason is UniEval is only trained for turn-level evaluation and it failed to transfer to dialogue-level evaluation.

6 Conclusion

In this work, we present **HOKIEBOT**, an open-domain chatbot aimed at addressing three essential challenges in dialogue systems for the Alexa SocialBot Challenge. Specifically, our system adapts to diverse conversational purposes, leverages effective memory management to facilitate consistent and user-preferred responses over long-term interactions, and incorporates a novel automatic evaluation mechanism for machine-generated dialogue responses. In future work, we plan to further investigate how to leverage explicit alignment with diverse human preferences to tackle fundamental challenges in text evaluation, such as alleviating the intrinsic biases of LLM-based evaluators.

References

Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. Keep me updated! memory management in long-term conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3769–3787, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.findings-emnlp.276`.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

Ssu Chiu, Maolin Li, Yen-Ting Lin, and Yun-Nung Chen. Salesbot: Transitioning from chit-chat to task-oriented dialogues, 2022.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416, 2022. doi: 10.48550/arXiv.2210.11416. URL `https://doi.org/10.48550/arXiv.2210.11416`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. MMDialog: A Large-scale Multi-turn Dialogue Dataset Towards Multi-modal Open-domain Conversation. *arXiv*, 2022. doi: 10.48550/arxiv.2211.05719.

Mauajama Firdaus, Nidhi Thakur, and Asif Ekbal. Aspect-Aware Response Generation for Multimodal Dialogue System. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(2):1–33, 2021. ISSN 2157-6904. doi: 10.1145/3430752.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. *CoRR*, abs/2302.04166, 2023. doi: 10.48550/arXiv.2302.04166. URL https://doi.org/10.48550/arXiv.2302.04166.

Rahul Goel, Shachi Paul, and Dilek Hakkani-Tür. Hyst: A hybrid approach for flexible and accurate dialogue state tracking. In *Interspeech 2019*, 2019. URL https://www.amazon.science/publications/hyst-a-hybrid-approach-for-flexible-and-accurate-dialogue-state-tracking.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anushree Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. Topical-chat: Towards knowledge-grounded open-domain conversations. In *Interspeech 2019*, 2019. URL https://www.amazon.science/publications/topical-chat-towards-knowledge-grounded-open-domain-conversations.

Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey P Bigham. Instructdial: improving zero and few-shot generalization in dialogue through instruction tuning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 505–525, 2022a.

Prakhar Gupta, Yang Liu, Di Jin, Behnam Hedayatnia, Spandana Gella, Sijia Liu, Patrick Lange, Julia Hirschberg, and Dilek Hakkani-Tur. Dialguide: Aligning dialogue model behavior with developer guidelines. *CoRR*, abs/2212.10557, 2022b. doi: 10.48550/arXiv.2212.10557. URL https://doi.org/10.48550/arXiv.2212.10557.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. Efficient natural language response suggestion for smart reply, 2017.

Jiun-Hao Jhan, Chao-Peng Liu, Shyh-Kang Jeng, and Hung-Yi Lee. Cheerbots: Chatbots toward empathy and emotionusing reinforcement learning, 2021.

Michael Johnston, Cris Flagg, Anna Gottardi, Sattvik Sahai, Yao Lu, Samyuth Sagi, Luke Dai, Prasoon Goyal, Behnam Hedayatnia, Lucy Hu, Di Jin, Patrick Lange, Shaohua Liu, Sijia Liu, Daniel Pressel, Hangjie Shi, Zhejia Yang, Chao Zhang, Desheng Zhang, Leslie Ball, Kate Bland, Shui Hu, Osman Ipek, James Jeun, Heather Rocker, Lavina Vaz, Akshaya Iyengar, Yang Liu, Arindam Mandal, Dilek Hakkani-Tür, and Reza Ghanadan. Advancing open domain dialog: The fifth alexa prize socialbot grand challenge. In *Alexa Prize SocialBot Grand Challenge 5 Proceedings*, 2023. URL https://www.amazon.science/publications/advancing-open-domain-dialog-the-fifth-alexa-prize-socialbot-grand-challenge.

Candace Kamm. User interfaces for voice applications. *Proceedings of the National Academy of Sciences*, 92 (22):10031–10037, 1995.

Chandra Khatri, Behnam Hedayatnia, Anu Venkatesh, Jeff Nunn, Yi Pan, Qing Liu, Han Song, Anna Gottardi, Sanjeev Kwatra, Sanju Pancholi, et al. Advancing the state of the art in open domain dialog systems through the alexa prize. *arXiv preprint arXiv:1812.10757*, 2018.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal inputs and outputs. 2023.

Ben Krause, Marco Damonte, Mihai Dobre, Daniel Duma, Joachim Fainberg, Federico Fancellu, Emmanuel Kahembwe, Jianpeng Cheng, and Bonnie Webber. Edina: Building an open domain socialbot with self-dialogues, 2017.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Minqian Liu and Lifu Huang. Teamwork is not always good: An empirical study of classifier drift in class-incremental information extraction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2241–2257, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-acl.141.

Minqian Liu, Shiyu Chang, and Lifu Huang. Incremental prompting: Episodic memory prompt for lifelong event detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2157–2165, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.189.

Minqian Liu, Ying Shen, Barry Menglong Yao, Sijia Wang, Jingyuan Qi, Zhiyang Xu, and Lifu Huang. Knowledgebot: Improving assistive robot for task completion and live interaction via neuro-symbolic reasoning. 2023a.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using GPT-4 with better human alignment. *CoRR*, abs/2303.16634, 2023b. doi: 10.48550/arXiv.2303.16634. URL https://doi.org/10.48550/arXiv.2303.16634.

Andrea Madotto, Zhaojiang Lin, Zhenpeng Zhou, Seungwhan Moon, Paul Crook, Bing Liu, Zhou Yu, Eunjoon Cho, Pascale Fung, and Zhiguang Wang. Continual learning in task-oriented dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7452–7467, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.590.

Shikib Mehri and Maxine Eskenazi. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting, July 2020a. Association for Computational Linguistics. URL https://aclanthology.org/2020.sigdial-1.28.

Shikib Mehri and Maxine Eskenazi. Usr: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, 2020b.

OpenAI. Gpt-4 technical report, 2023.

Yan Pan, Mingyang Ma, Bernhard Pflugfelder, and Georg Groh. How to build robust faq chatbot with controllable question generator?, 2021.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019.

Jingyuan Qi, Zhiyang Xu, Ying Shen, Minqian Liu, Di Jin, Qifan Wang, and Lifu Huang. The art of socratic questioning: Zero-shot multimodal reasoning with recursive thinking and self-questioning. *arXiv preprint arXiv:2305.14999*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, 2021.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *CoRR*, abs/2208.03188, 2022. doi: 10.48550/arXiv.2208.03188. URL https://doi.org/10.48550/arXiv.2208.03188.

Saleh Soltan, Shankar Ananthakrishnan, Jack FitzGerald, Rahul Gupta, Wael Hamza, Haidar Khan, Charith Peris, Stephen Rawls, Andy Rosenbaum, Anna Rumshisky, et al. Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model. *arXiv preprint arXiv:2208.01448*, 2022.

Kai Sun, Seungwhan Moon, Paul Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. Adding chit-chat to enhance task-oriented dialogues, 2021.

Qingfeng Sun, Yujing Wang, Can Xu, Kai Zheng, Yaming Yang, Huang Hu, Fei Xu, Jessica Zhang, Xiubo Geng, and Daxin Jiang. Multimodal Dialogue Response Generation. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2854–2866, 2022a. doi: 10.18653/v1/2022.acl-long.204.

Qingfeng Sun, Can Xu, Huang Hu, Yujing Wang, Jian Miao, Xiubo Geng, Yining Chen, Fei Xu, and Daxin Jiang. Stylized knowledge-grounded dialogue generation via disentangled template rewriting, 2022b.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. PARADISE: A framework for evaluating spoken dialogue agents. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280, Madrid, Spain, July 1997. Association for Computational Linguistics. doi: 10.3115/976909.979652. URL `https://aclanthology.org/P97-1035`.

Sijia Wang, Alexander Hanbo Li, Henry Zhu, Sheng Zhang, Chung-Wei Hang, Pramuditha Perera, Jie Ma, William Wang, Zhiguo Wang, Vittorio Castelli, et al. Benchmarking diverse-modal entity linking with generative models. *arXiv preprint arXiv:2305.17337*, 2023.

Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. Wikidiverse: A multimodal entity linking dataset with diversified contextual topics and entity types. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4785–4797, 2022.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=gEZrGCozdqR`.

Yu Wu, Wei Wu, Chen Xing, Can Xu, Zhoujun Li, and Ming Zhou. A sequential matching framework for multi-turn response selection in retrieval-based chatbots. *Comput. Linguist.*, 45(1):163–197, mar 2019. ISSN 0891-2017. doi: 10.1162/coli_a_00345. URL `https://doi.org/10.1162/coli_a_00345`.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*, 2023a.

Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland, May 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.356. URL `https://aclanthology.org/2022.acl-long.356`.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. Long time no see! open-domain conversation with long-term persona memory. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2639–2650. Association for Computational Linguistics, 2022b. doi: 10.18653/v1/2022.findings-acl.207. URL `https://doi.org/10.18653/v1/2022.findings-acl.207`.

Zhiyang Xu, Ying Shen, and Lifu Huang. MultiInstruct: Improving multi-modal zero-shot learning via instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11445–11465, Toronto, Canada, July 2023b. Association for Computational Linguistics. URL `https://aclanthology.org/2023.acl-long.641`.

Barry Menglong Yao, Yu Chen, Qifan Wang, Sijia Wang, Minqian Liu, Zhiyang Xu, Licheng Yu, and Lifu Huang. Ameli: Enhancing multimodal entity linking with fine-grained attributes. *arXiv preprint arXiv:2305.14725*, 2023.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277, 2021.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. DynaEval: Unifying turn and dialogue level evaluation. 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 5676–5689, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.441. URL `https://aclanthology.org/2021.acl-long.441`.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SkeHuCVFDr`.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 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*, pages 270–278, 2020.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.emnlp-main.131`.