This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3409377

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.0322000

Flowlogue: A Novel Framework for Synthetic Dialogue Generation with Structured Flow from Text Passages

YONGIL KIM[1,3], YERIN HWANG[2,3], HYUNKYUNG BAE[4], TAEGWAN KANG[4], and KYOMIN JUNG[1,2,3,5]

[1]Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, South Korea
[2]Automation and Systems Research Institute, Seoul National University, Seoul 08826, South Korea
[3]Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul 08826, South Korea
[4]LG AI Research, Seoul 07796, South Korea
[5]SNU-LG AI Research Center, Seoul 08826, South Korea

Corresponding author: Kyomin Jung (e-mail: kjung@snu.ac.kr).

ABSTRACT Dialogue systems play a pivotal role in domains ranging from customer service to virtual assistance and education, using natural language to deliver information and resolve inquiries. Integrating Large Language Models (LLMs) has significantly boosted their capabilities and applications, underscoring their potential to facilitate more nuanced human-computer interactions. Despite these advances, a significant challenge persists in curated dialogue data scarcity, especially in Conversational Question Answering (ConvQA) systems that require domain-specific information. Traditional Passage to Dialogue (P2D) methods attempt to mitigate this by converting textual passages into dialogue form but often need help with issues such as unnatural responses and information redundancy due to the direct use of passage sentences as dialogue answers. To overcome these limitations, we introduce Flowlogue, a novel ConvQA framework that enhances dialogue generation by merging related sentences within passages to maintain natural flow and coherence. This approach leverages LLMs to generate questions and contextually relevant answers based on newly formed dialogue flows, significantly improving the quality and relevance of dialogues compared to existing P2D methods. Our experimental results, validated through reference-free metrics and GPT-4 evaluations, confirm that Flowlogue produces superior dialogues, establishing a robust framework for generating natural, high-quality ConvQA dialogues that effectively harness the depth and nuance of human conversations.

INDEX TERMS Conversational question answering, data generation framework, dialogue system, synthetic dialogue generation

I. INTRODUCTION

DIALOGUE systems [1, 2], also known as conversational agents or chatbots, are designed to exchange conversations in natural language, aiming to provide users with relevant information or resolve their inquiries [3]. These systems find applications across various domains, ranging from customer service bots to virtual assistants like Siri, Alexa, and Google Assistant, as well as therapeutic bots and educational tools [4, 5, 6]. In particular, the emergence of Large Language Models (LLMs) [7, 8] has notably drawn considerable attention to dialogue systems from both the industry and academic sectors [9, 10]. This development highlights these systems' increasing importance and potential in enabling more natural and effective human-computer interactions.

However, one of the most significant challenges in dialogue systems is the data scarcity problem [11], wherein, despite the abundance of human-written textual passages available online, there is a need for well-curated dialogue data. In particular, the data scarcity issue becomes even more pronounced in the context of information-seeking Conversational Question Answering (ConvQA) [12, 13], where the objective is to convey specific topic information to humans. This is primarily because the generation of ConvQA data that includes accurate information about a domain necessitates the direct involvement of domain experts in the data annotation process [14].

In an effort to resolve this issue, Passage to Dialogue (P2D) methods [15, 16] have been introduced to create dialogue data from passages. These methodologies involve segmenting a passage into sentences, utilizing each sentence directly as an

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3409377

TABLE 1. Error cases in existing P2D methods. Case 1: Starting a sentence with conjunctions like "however" often results in questions that yield unnatural-sounding answers. Case 2: When all the details are included in the answer, it can cause information redundancy, making even the most well-formulated questions seem unnatural. Case 3: Simply breaking a text into individual sentences without considering the links between them overlooks a crucial element of coherence necessary for constructing dialogues.

Error Case	Passage	Dialogue Generated by P2D methods
Case 1	... She then took work as a contributing editor. However, in November 1937 she was arrested by the NKVD and sentenced to three and a half years in a concentration camp. ...	Q: What type of publications did Roberta Gropper contribute to as a contributing editor after her husband's passing in May 1935? A: She then took work as a contributing editor. Q: What led to Roberta Gropper's arrest by the NKVD in November 1937? A: **However**, in November 1937 she was arrested by the NKVD and sentenced to three and a half years in a concentration camp.
Case 2	... People who took part in the Auckland digital trial using terrestrial H.262 receivers needed to change their receivers to more expensive H.264 models in order to receive terrestrial Freeview. ...	Q: Did people who took part in the Auckland digital trial using terrestrial H.262 receivers have to change their receivers to more expensive H.264 models in order to receive terrestrial Freeview? A: People who took part in the Auckland digital trial using terrestrial H.262 receivers needed to change their receivers to more expensive H.264 models in order to receive terrestrial Freeview.
Case 3	... Blatty said, "I considered it a comic novel, but a great deal of philosophy and theology crept into it. But the farcical elements outweighed the serious elements." ...	Q: What was Blatty's intention when he adapted the novel into a screenplay and planned for it to be filmed by William Friedkin? A: 'Blatty said, "I considered it a comic novel, but a great deal of philosophy and theology crept into it. Q: What did Blatty say about the balance between comedy and seriousness in his novel? A: But the farcical elements outweighed the serious elements."

answer in dialogue, and then employing a question generator to craft suitable questions for each answer, thereby enabling the automatic creation of ConvQA data. These P2D methods ensure consistency and a natural dialogue flow by directly translating the flow of the passage into the flow of the conversation, offering significant advantages in maintaining the coherence and fluidity of dialogues.

Nevertheless, existing P2D methods, which segment passages into sentence units and directly use these fixed answers as responses in dialogue, face several limitations. Firstly, when a sentence begins with conjunctions such as "however," crafting any question leads to an answer that does not sound natural (Case 1 in Table I). Furthermore, in natural human conversations, information is appropriately distributed between the question and answer. However, when using a passage directly as an answer, all the information is contained within the answer. This often leads to information redundancy, even with well-crafted questions, resulting in unnaturalness (Case 2 in Table I). Moreover, the approach of merely dividing a passage into sentence units fails to consider the relationships or connections between sentences, overlooking an essential aspect of coherence in dialogue construction (Case 3 in Table I).

To address these issues, we propose the novel ConvQA generation framework, Flowlogue, which involves appropriately merging sentences within a passage to utilize them in the dialogue flow. Initially, the passage is segmented into sentences, then the similarity between each pair of adjacent sentences is calculated. The two sentences with the highest similarity are merged into one, and this process is repeated until either the target number of dialogue flows is achieved or only sentences with similarity below a certain threshold remain. Upon completing the dialogue flow, an LLM is utilized as a question generator to create questions that are suited to each dialogue flow, thereby finalizing the draft of the dialogue. Subsequently, using the dialogue history, target question, and passage, a contextually relevant answer is generated for the specific question, completing the creation of high-quality and natural information-seeking ConvQA data. This method represents the first approach in synthetic dialogue generation that creates dialogue flows from passages, offering a way to maintain dialogue consistency while generating natural dialogues.

Through the experiments, we quantitatively and qualitatively demonstrate that dialogues generated by Flowlogue are of higher quality and more contextually relevant compared to those produced by existing P2D methods. Firstly, we employ various reference-free metrics [17, 18] for automatic evaluation to prove that our methodology generates superior dialogues compared to existing methods. Subsequently, through GPT-4 evaluation [19, 20], we demonstrate that dialogues created using Flowlogue receive higher scores across multiple criteria than those produced by the conventional method. Additionally, we validate the efficacy of the proposed Flowlogue methodology through analysis with an ablation study and case study.

The contributions of this research are as follows:

- We introduce the Flowlogue framework, pioneering the use of passages as dialogue flows for synthesizing natural conversational question-answering data.
- We advance the methodology in synthetic dialogue gen-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3409377

eration by implementing a novel sentence-merging strategy that enhances dialogue coherence and effectively addresses prior limitations.
- We demonstrate quantitatively and qualitatively that dialogues generated by Flowlogue surpass the quality of those produced using existing P2D methods.

The remaining parts of our paper are structured into six sections, aimed at offering a comprehensive overview and analysis of our study. Section II embarks on a brief review of previous works, specifically focusing on passage-to-dialogue methods and reference-free dialogue metrics, laying the groundwork for our research. In Section III, we introduce our dialogue generation framework, outlining its components. We present the experimental results in Section IV. Moving forward, Section V encompasses a diverse array of analyses, including an ablation study and a case study, which further elucidate the nuances of our findings. Finally, Section VI wraps up our paper with a conclusion, highlighting the key takeaways and acknowledging the limitations of our study.

II. RELATED WORKS

A. CONVERSATIONAL QUESTION ANSWERING (CONVQA)

Conversational question-answering (ConvQA) [21] aims to engage in dialogue with users to provide information. The development and refinement of ConvQA systems have garnered significant interest due to their wide range of applications in customer service, virtual assistance, educational platforms, and more [22, 23, 24]. However, the effectiveness of a ConvQA system heavily relies on the quality and scope of its training data [12]. The training datasets must be rich in domain-specific knowledge and constructed in a way that captures the nuances of human dialogue. This often requires substantial human effort in curating and annotating data, where domain experts contribute to ensuring the relevance and accuracy of the information. Consequently, existing ConvQA datasets such as CoQA [14], CSQA [25], and ConvQuestions [26] have been created with substantial human efforts. In this research, we introduce a method aimed at automating the creation of ConvQA datasets by utilizing textual sources.

B. PASSAGE TO DIALOGUE (P2D)

Recently, Passage to Dialogue (P2D) frameworks [15, 16] have gained attention as a means to address the data scarcity issue in ConvQA. These frameworks are designed to facilitate the efficient creation of dialogues from textual content without compromising the integrity of the information. As shown in Figure 1, they work by dividing text passages into sentences, which are treated as "answers," and then utilizing question generation models trained on specific tasks to produce corresponding "questions" for each "answer." While there is a limited amount of expert-generated conversational data available online, there is an abundance of well-curated textual data authored by experts. This discrepancy highlights the potential of such approaches to solve the data scarcity problem in ConvQA.

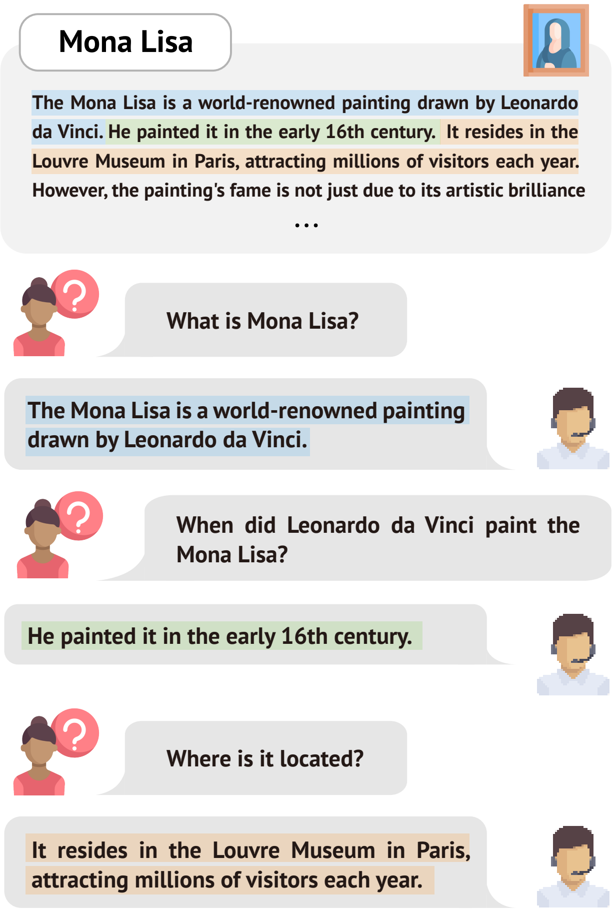
Mona Lisa
The Mona Lisa is a world-renowned painting drawn by Leonardo da Vinci. He painted it in the early 16th century. It resides in the Louvre Museum in Paris, attracting millions of visitors each year. However, the painting's fame is not just due to its artistic brilliance
...
What is Mona Lisa?
The Mona Lisa is a world-renowned painting drawn by Leonardo da Vinci.
When did Leonardo da Vinci paint the Mona Lisa?
He painted it in the early 16th century.
Where is it located?
It resides in the Louvre Museum in Paris, attracting millions of visitors each year.

FIGURE 1. Passage to Dialogue (P2D) Framework. This framework operates by segmenting text passages into sentences, labeled as "answers," and then uses models trained in question generation for specific tasks to create matching "questions" for each "answer."

The concept of a P2D framework was initially introduced with a model featuring a question generator named Dialog Inpainter [15], developed through a dialogue reconstruction task. This task involves masking parts of a dialogue's utterances and training the model to reconstruct these masked sections. Furthermore, an advanced P2D framework named Dialogizer [16] has been proposed, which incorporates additional tasks enabling the learning of sentence-level alignment within dialogues and thereby enhancing the generation of ConvQA data. This enhanced P2D framework aims to improve the contextual relevance of the generated dialogues, addressing one of the key challenges in the field.

These methods can transform specialized passages into dialogues without any loss of information, even without the participation of a domain expert. However, these methods, while effective, often result in a disjointed narrative. For example, the use of sentences starting with conjunctions like ’however’ as standalone responses can feel artificial. Additionally, this approach can lead to repetitive dialogue sequences, particularly when the questions and answers relate to closely linked content spread across multiple sentences. Our

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3409377

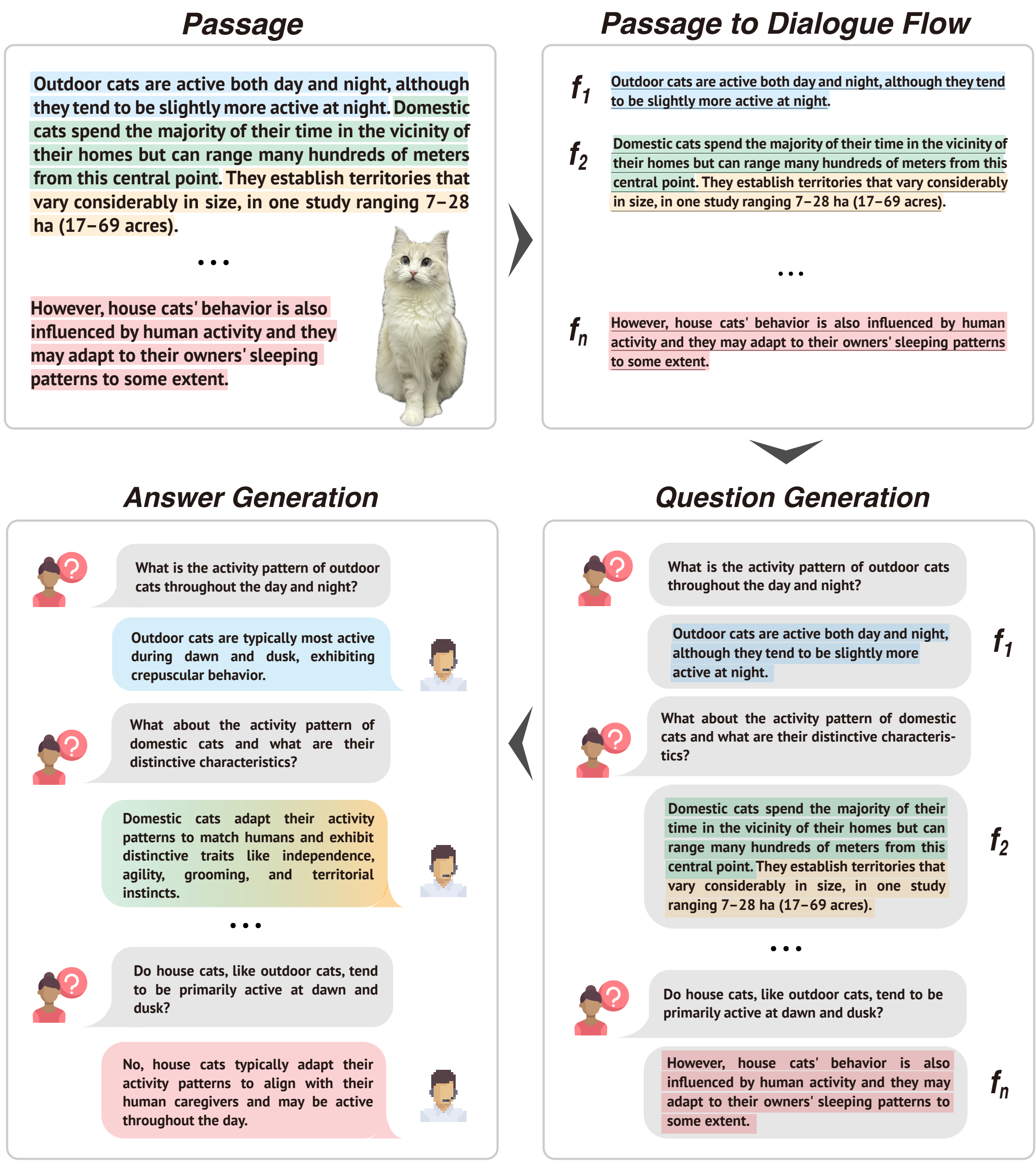
Passage
Outdoor cats are active both day and night, although they tend to be slightly more active at night. Domestic cats spend the majority of their time in the vicinity of their homes but can range many hundreds of meters from this central point. They establish territories that vary considerably in size, in one study ranging 7–28 ha (17–69 acres).
• • •
However, house cats' behavior is also influenced by human activity and they may adapt to their owners' sleeping patterns to some extent.
Passage to Dialogue Flow
f1
Outdoor cats are active both day and night, although they tend to be slightly more active at night.
f2
Domestic cats spend the majority of their time in the vicinity of their homes but can range many hundreds of meters from this central point. They establish territories that vary considerably in size, in one study ranging 7–28 ha (17–69 acres).
• • •
fn
However, house cats' behavior is also influenced by human activity and they may adapt to their owners' sleeping patterns to some extent.
Answer Generation
What is the activity pattern of outdoor cats throughout the day and night?
Outdoor cats are typically most active during dawn and dusk, exhibiting crepuscular behavior.
What about the activity pattern of domestic cats and what are their distinctive characteristics?
Domestic cats adapt their activity patterns to match humans and exhibit distinctive traits like independence, agility, grooming, and territorial instincts.
• • •
Do house cats, like outdoor cats, tend to be primarily active at dawn and dusk?
No, house cats typically adapt their activity patterns to align with their human caregivers and may be active throughout the day.
Question Generation
What is the activity pattern of outdoor cats throughout the day and night?
Outdoor cats are active both day and night, although they tend to be slightly more active at night.
f1
What about the activity pattern of domestic cats and what are their distinctive characteristics?
Domestic cats spend the majority of their time in the vicinity of their homes but can range many hundreds of meters from this central point. They establish territories that vary considerably in size, in one study ranging 7–28 ha (17–69 acres).
f2
• • •
Do house cats, like outdoor cats, tend to be primarily active at dawn and dusk?
However, house cats' behavior is also influenced by human activity and they may adapt to their owners' sleeping patterns to some extent.
fn

FIGURE 2. Overall Flowlogue Framework. Passage to Dialogue Flow: Sentences in passages are grouped into dialogue flows based on content similarity, ensuring coherence with consecutive. Question Generation: Questions are crafted to align with the natural speech patterns and the context of the dialogue flow, focusing on its key points. Answer Generation: Answers are then developed using the dialogue flow as a reference, ensuring they respond precisely and thoroughly to the posed questions, thus maintaining a seamless and informative dialogue.

ConvQA generation framework addresses these issues, providing a more cohesive and natural dialogue experience. This research presents a novel ConvQA generation framework that employs continuous sentence spans to create questions within dialogue flows. Subsequently, answers are generated using the questions and dialogue flows, thereby facilitating natural conversations.

C. REFERENCE-FREE DIALOGUE METRICS

Due to the one-to-many nature of dialogue and question generation tasks [27], traditional reference-based metrics for evaluating natural language generation, such as BLEU [28] and ROUGE [29], often fail to align well with human judgments. This discrepancy has led to a growing emphasis on reference-free evaluation metrics [30, 31], which exhibit stronger correlations with human assessments. In this study, we employ a range of reference-free metrics [17, 18] to gauge the quality of the generated ConvQA datasets, demonstrating their superior quality. We use metrics to measure the context coherence within the dialogue and the relevance between the question and context.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3409377

III. FLOWLOGUE

We propose the Flowlogue framework, a novel ConvQA generation framework that utilizes passages as dialogue flows for generating natural ConvQA data. The framework uses continuous sentence spans to map dialogue flows, generating questions that facilitate natural conversations. It then produces answers based on these questions and flows, ensuring information is retained. In the following subsections, we introduce the three stages: passage to dialogue flow, question generation, and answer generation, detailing each step by step. Figure 2 illustrates the overall framework.

A. PASSAGE TO DIALOGUE FLOW

In typical human interactions, the dissemination of information is effectively balanced between questions and responses. However, in traditional Passage-to-Dialogue methods [15, 16], the conventional approach involves using every sentence in a passage as a response to a conversation. This practice presents several limitations. Firstly, it often leads to redundancies in dialogue turns, as adjacent sentences in a passage frequently share similar topics or information. As a result, corresponding dialogue turns may contain repetitive content.

Flowlogue addresses the issue of redundancy by dividing a passage into variable segments, each comprising either a single or multiple consecutive sentences. In detail, we calculate the sentence similarity [32] for all consecutive pairs within the passage. For a passage defined as $\{s_1, s_2, ..., s_n\}$, we assess the similarity scores for each consecutive pair, $\{sim(s_1, s_2), sim(s_2, s_3), ..., sim(s_{n-1}, s_n)\}$, where sim represents the sentence similarity score. We then merge the pair with the highest similarity score into a dialogue flow. During this process, we maintain a threshold to ensure the similarity score remains above a specified minimum. This merging process is repeated until the threshold is consistently met. To prevent an excessive reduction in the number of turns in the dialogue, we set a minimum length (*min_length*) to guarantee a specified number of dialogue turns. The pseudo-code for the procedure that transforms a passage into a dialogue flow is presented in Algorithm 1.

B. QUESTION GENERATION

Upon transforming the sentence spans from the passage into a Dialog Flow, questions are subsequently formulated with this dialog flow serving as the response, thus establishing a dialogic structure. This methodology enables the construction of an info-seeking dialog format.

We use the LLMs [33] as the question generator in Flowlogue. This choice is grounded on insights obtained from analyzing datasets produced by different P2D models and LLMs, which suggest that LLMs potentially generate questions for dialogue turns more efficiently. In our approach, we employ LLMs as question generators to carry out a dialogue reconstruction task, filling in [BLANK] without the need for a specific prompt. The prompt used in the LLM is as shown in Table 2.

Algorithm 1 Passage to Dialogue flow

```
Input: Original passage
Output: Dialogue flow
procedure Passage to Dialogue Flow
    for each sentence in the original passage do
        Calculate similarity scores for all consecutive sentence pairs
        Find the index of the maximum similarity score
        while number of scores $\geq$ $min\_length$ and maximum score $\geq$ threshold do
            Combine the two sentences at the index of maximum similarity
            Update the dialogue list by replacing the two sentences with their combination
            Recalculate the similarity scores
            Update the index of the maximum similarity score
        end while
        Add the updated dialogue to the dialogue flow
    end for
    return Dialogue flow
end procedure
```

TABLE 2. The template of the prompt used for question generation process in Flowlogue.

```
You are an automatic assistant that generates appropriate question based on the predefined answer. Generate a single question that is most suitable for the given dialogue history and target answer.

Please fill in only [BLANK] in the next dialogue.

START
A: {question 1}
B: {answer 1}
...
A: [BLANK]
B: {answer t}
END
```

C. ANSWER GENERATION

The P2D method effectively creates information-seeking dialogues [34] from general passages. However, most sentences in passages are not designed as responses to specific questions, making their use in ConvQA systems feel unnatural. Moreover, sentences that start with conjunctions typically serve to connect thoughts, which makes them awkward as standalone responses in a dialogue. As a result, when P2D methods segment passages solely into discrete sentence units, they tend to produce dialogues that lack natural flow. This approach overlooks crucial aspects of coherence, essential for

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3409377

creating dialogues that realistically mimic human interaction.

Flowlogue addresses these issues by regenerating natural answers based on the questions created. During the answer generation process, the system uses the existing dialogue flow as a contextual guide to shape the response. Since the dialogue flow is provided as input, it allows for forming very natural and well-connected dialogues without losing information. The answer-generation process is implemented using an LLM, similar to the question-generation process.

IV. EXPERIMENTS

In this section, we provide empirical evidence demonstrating the Flowlogue framework's ability to generate high-quality dialogues quantitatively and qualitatively. We use P2D methods and an LLM as baseline question generators, comparing these to Flowlogue using various reference-free dialogue metrics. Additionally, we enhance our qualitative analysis by using GPT-4 to assess dialogue quality.

A. EXPERIMENTAL SETTINGS

We perform comparative evaluations by generating 1,000 multi-turn dialogues with the Wikipedia dataset. Each generated dialogue is then assessed to evaluate different methodologies. In Flowlogue, we use the GPT-3.5-turbo model [1] for both question and answer generation phases. The generation processes employ a structured prompt based on the template outlined in Section III, ensuring consistency across evaluations. We also evaluate the Flowlogue model, which only proceeds to question generation as a baseline model. Subsequently, we compare these baseline methods with the complete Flowlogue framework that includes answer generation. This comparison aims to highlight the enhanced capabilities of Flowlogue, particularly its ability to generate coherent and contextually rich answers.

B. BASELINE METHODS

We carry out experiments to compare our method with three P2D methods for generating dialogues: Dialog Inpainter [15], Dialogizer [16], and the LLMs. The Dialog Inpainter and Dialogizer are adapted to meet specific framework requirements to make the comparison fair. Both models are built on the T5-base [35] architecture and trained by four different datasets: Task Masker [36], Daily Dialog [37], OR-QUAC [38], and QReCC [39]. For the LLM, we utilize GPT-3.5-turbo, augmenting it with specific instructions to effectively handle topic changes during the question generation phase.

C. FLOWLOGUE VARIANTS

The Flowlogue framework incorporates sentence similarity metrics to facilitate the transformation of passages into dialogue flows. We employ three distinct sentence similarity metrics: BERTScore [40], which is based on BERT; LeallaScore [41], a proprietary metric; and GTEScore [42], which is derived from Sentence Transformers [43]. Each metric is selected for its unique approach to measuring textual similarities. The application of these metrics on consecutive sentence pairs in Wikipedia passages yielded results that are graphically represented in Figure 3.

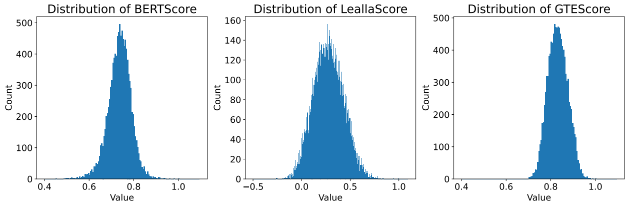

FIGURE 3. Histograms displaying the distribution of sentence similarity scores for three different metrics: BERTScore, LellaScore and GTEScore.

Following the data presented in Figure 3, we have empirically determined optimal thresholds to enhance the efficacy of our framework. Specifically, we have established thresholds of 0.75 for BERTScore, 0.25 for LellaScore, and 0.85 for GTEScore, each tailored to the unique attributes of the respective metrics. Additionally, we have instituted a requirement that all frameworks produce dialogues with a minimum of seven sentences, thereby ensuring a substantive length in the dialogues generated.

D. METRICS

To demonstrate the performance of the Flowlogue framework as a dialogue generation framework, we utilize various reference-free metrics to thoroughly assess dialogues or generated questions, allowing for a quantitative analysis of Flowlogue's efficacy. Central to our evaluation methodology is USR-DR [30], a distinguished reference-free dialogue metric that examines dialogues for context coherence, engagement, and the effective use of knowledge. This metric includes USR-DR(c), focusing on dialogue evaluation through historical and factual inputs, and USR-DR(f), which prioritizes factual context in its assessment. In addition, a GPT-2-based metric [44] scrutinizes dialogues for utterance coherence. Moreover, RQUGE [17] and QRelScore [18] offer insights into question answerability within context and context-aware question generation capabilities without requiring additional training or human oversight. QRelScore further bifurcates into QRelScore$_{\text{LRM}}$, which delves into complex reasoning via word-level similarity, and QRelScore$_{\text{GRG}}$, aimed at verifying factual accuracy through the confidence in generating contextually pertinent content.

E. MAIN RESULTS

Table 3 shows the comparative experiment results between three existing P2D Methods and our Method. The frameworks marked with dagger symbol (†) utilize BERTScore, those marked with an asterisk symbol (*) use LellaScore, and the unmarked variants of the Flowlogue framework employ GTEScore as the sentence similarity metric. Initially, we test models that generate only the question using dialogue flow while retaining the original answer from the passage.

[1] https://platform.openai.com/docs/models/gpt-3-5-turbo

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3409377

TABLE 3. Main Experiments Results. Flowlogue[†] uses BERTScore, Flowlogue* uses LellaScore, and Flowlogue uses GTEScore as the sentence similarity metric to create dialogue flows. When merely aggregating consecutive sentence pairs to serve as answers, the resultant performance is inferior to that achieved by utilizing a single sentence segment as an answer. Conversely, producing natural answers through answer generation consistently yields superior performance across the board.

Framework	Question Generation	Answer Generation	USR-DR (f)	USR-DR (c)	GPT-2	RQUGE	$\text{QRelScore}_{\text{LRM}}$	$\text{QRelScore}_{\text{GRG}}$
P2D	*Dialog Inpainter*	-	0.9615	0.7227	0.5125	3.1255	0.4887	0.4808
	Dializiger	-	0.9641	0.7883	0.5386	3.2511	0.5044	0.4852
	GPT-3.5-turbo	-	0.9856	0.8960	0.5739	3.2923	0.5369	0.5305
Flowlogue[†]	*GPT-3.5-turbo*	-	0.8684	0.7931	0.5086	2.9183	0.4755	0.4702
	GPT-3.5-turbo	*GPT-3.5-turbo*	0.9877	0.9123	0.5915	4.0387	0.5617	0.5584
Flowlogue*	*GPT-3.5-turbo*	-	0.8661	0.7871	0.5038	2.8931	0.4709	0.4662
	GPT-3.5-turbo	*GPT-3.5-turbo*	0.9874	**0.9134**	0.5921	4.0411	0.5598	0.5578
Flowlogue	*GPT-3.5-turbo*	-	0.8789	0.7978	0.5123	2.9314	0.4781	0.4733
	GPT-3.5-turbo	*GPT-3.5-turbo*	**0.9878**	0.9133	**0.5933**	**4.0893**	**0.5628**	**0.5599**

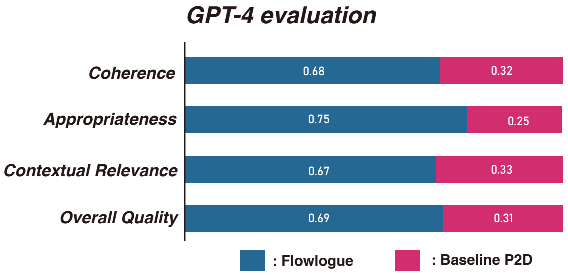

FIGURE 4. GPT-4 Evaluation. We demonstrate the superiority of the Flowlogue framework over the baseline P2D method through a win/lose comparison. Flowlogue consistently receives excellent evaluations.

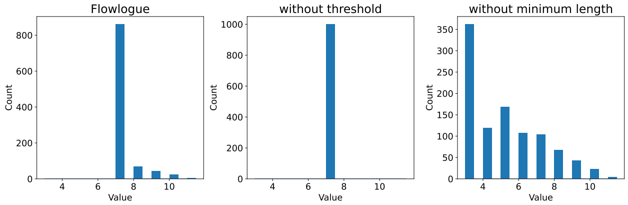

FIGURE 5. Histograms displaying the average length of dialogues for ablation study.

Methods that simply cluster sentences in the passage to create questions underperform compared to existing P2D methods that craft answers from individual sentences, primarily due to excessive answer lengths that hamper effective relevance measurement. However, when answers are naturally generated using our proposed dialogue flow methodology, the models demonstrate enhanced performance across all evaluation metrics. The performance remains consistent across all sentence similarity metrics. Our proposed framework, which integrates both question and answer generation, consistently demonstrates its capability to create highly natural conversational QA dialogues.

F. GPT-4 EVALUATIONS

Based on findings from [45] that demonstrate strong LLM judges like GPT-4 can match both controlled and crowd-sourced human preferences well, we qualitatively evaluate the Flowlogue framework using GPT-4 as an evaluator. Our evaluations are based on four critical criteria—coherence, appropriateness, contextual relevance, and overall quality—as defined by the characteristics of the ConvQA and dialogue [46]. Coherence measures the logical connectivity and fluidity of the conversation across the question, context, and answer; appropriateness assesses whether the question, context and answer fit conversational norms and expectations within the dialogue; contextual relevance evaluates how the question aligns with the context and answer; and overall quality assesses the coherence and clarity of the dialogue interaction.

We confirm the superiority of the Flowlogue framework through a win/lose comparison with the baseline P2D method. As shown in Figure 4, Flowlogue receives impressive evaluations across all criteria. It validates that the questions and answers generated through Flowlogue are composed of very natural and high-quality sentences compared to baseline P2D method.

V. ANALYSIS

In this section, we conduct further analysis of the Flowlogue framework. Initially, we demonstrate the necessity of the elements used in forming dialogue flows in Flowlogue—threshold and minimum length—through an ablation study. Subsequently, we perform a further qualitative survey via a step-by-step case study of the Flowlogue framework.

A. ABLATION STUDY

We conduct an ablation study on the threshold (τ) and minimum length (*min_length*) for forming dialogue flows. Our proposed framework, Flowlogue, utilizes both elements and employs LellaScore as the sentence similarity metric in this ablation study. As shown in Figure 5, without a threshold τ, all dialogues are truncated to the *min_length* 7, and without a *min_length*, the dialogues can spread down to fewer than seven turns. The results in Table 5 reveal that while the average dialogue length in Flowlogue is 7.248, with-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3409377

TABLE 4. Qualtitative Study. We specifically analyze how Flowlogue tackles three limitations identified in the existing P2D method.

Case	Dialogue Generated by P2D methods	Generated Dialogue by Flowlogue
Case 1	Q: What role does Robert California assign Gabe at the Sabre headquarters in the episode Trivia? A: Robert then informs Dwight that he cannot meet with him, but will have him meet with Bill, another executive, much to Dwight's frustration. Q: What does Robert California instruct Gabe to do regarding Dwight's pitch? A: **However**, Robert manipulates the situation by secretly calling Gabe and instructing him to not have Dwight speak with Bill either, but to have Gabe listen to Dwight's pitch and then reject him.	Q: What role does Robert California assign Gabe at the Sabre headquarters in the episode Trivia? A: Robert California assigns Gabe the role of listening to Dwight's pitch and then rejecting him at the Sabre headquarters in the episode Trivia.
Case 2	Q: Did people who took part in the Auckland digital trial using terrestrial H.262 receivers have to change their receivers to more expensive H.264 models in order to receive terrestrial Freeview? A: People who took part in the Auckland digital trial using terrestrial H.262 receivers needed to change their receivers to more expensive H.264 models in order to receive terrestrial Freeview.	Q: Did people who took part in the Auckland digital trial using terrestrial H.262 receivers have to change their receivers to more expensive H.264 models in order to receive terrestrial Freeview? A: Yes, they needed to change their receivers to more expensive H.264 models in order to receive terrestrial Freeview.
Case 3	Q: What prompted Keable's interest in the concept of creative evolution while exploring various philosophical and scientific works after leaving the Anglican church? A: He wrote, of the history of Christianity, "I can see creative evolution at work. Q: What new perspective did Keable gain from his exploration of creative evolution and its connection to Christianity after leaving the Anglican church? A: What is behind it, I don't know"	Q: What did Keable write about his view on the history of Christianity? A: He wrote, 'I can see creative evolution at work. What is behind it, I don't know.

TABLE 5. Ablation study.

	Avg. len	RQUGE	USR-DR	QRelScore	GPT-2
Flowlogue	7.248	4.0893	0.9878	0.5628	0.5933
- *thershold* τ	7.0	3.8875	0.9858	0.5542	0.5815
- *min_length*	4.998	3.8842	0.9803	0.5479	0.5793

out the threshold τ, it truncates to the minimum length of 7. When merging consecutive sentences to create dialogue flows, including those with lower similarity results in a slight performance decrease. Not specifying a *min_length* leads to a significant shortening of dialogue average length to 4.998, resulting in substantial information loss during the process of generating new answers based on the dialogue flow and, ultimately, a decrease in performance across all metrics. This confirms that setting appropriate thresholds and *min_length* is crucial for generating natural dialogues.

B. QUALITATIVE STUDY

In this section, we further explore the Flowlogue framework qualitatively through a step-by-step case study. Specifically, we investigate how Flowlogue addresses three points highlighted as limitations of the existing P2D method. Examples of dialogues generated using the proposed Flowlogue framework can be found in Table 4. Firstly, addressing the issue of unnatural answers when starting with conjunction, Flowlogue combines consecutive sentences containing conjunctions into a single natural response within the dialogue, ensuring no loss of information through both dialogue flow and answer generation stages. In the second scenario, where answers contain substantial information, redundancy arises with question content closely mirroring the answer. This redundancy is effectively eliminated by providing a natural response to the question during the answer generation step. Lastly, in the third case, where the P2D method segments responses into single sentences, resulting in disjointed quotes and awkwardness, Flowlogue seamlessly integrates quotes during the dialogue flow generation process, presenting them naturally as answers during answer generation.

VI. CONCLUSION

We identify a critical challenge in the field: the scarcity of high-quality, domain-specific dialogue data, particularly for Conversational Question Answering (ConvQA). Our paper notes that existing Passage to Dialogue (P2D) methods, which convert text passages into dialogue form, often lead to unnatural dialogue sequences due to their simplistic sentence-by-sentence conversion approach. To overcome these limitations, we propose a new ConvQA generation framework called Flowlogue. We innovate by merging sentences based on similarity, ensuring a more natural flow and coherence in dialogues. We employ an LLM to generate questions tailored to these merged sentence units, enabling the creation of more natural and contextually relevant dialogue data. We provide evidence through rigorous evaluations, including GPT-4 assessments, that dialogues generated by Flowlogue are of higher quality and more contextually appropriate compared to those created by existing P2D methods. This represents a significant step forward in synthetic dialogue generation, maintaining consistency and enhancing the naturalness of dialogues.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3409377

ACKNOWLEDGMENT

This work was supported by LG AI Research. This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University) & NO.2021-0-02068, Artificial Intelligence Innovation Hub (Artificial Intelligence Institute, Seoul National University)], the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2023, and the National Research Foundation of Korea (NRF) grant funded by the Korea government (No. 2021R1A2C2008855). K. Jung is with ASRI, Seoul National University, Korea.

REFERENCES

[1] S. Arora, K. Batra, and S. Singh, "Dialogue system: A brief review," *arXiv preprint arXiv:1306.4134*, 2013.

[2] J. Ni, et al., "Recent advances in deep learning based dialogue systems: A systematic survey," *Artificial Intelligence Review*, vol. 56, no. 4, pp. 3055-3155, 2023.

[3] T. D. Oesterreich, et al., "How can I help you? Design principles for task-oriented speech dialog systems in customer service," *Information Systems and e-Business Management*, vol. 21, no. 1, pp. 37-79, 2023.

[4] L. Brocki, et al., "Deep learning mental health dialogue system," in *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*, IEEE, 2023.

[5] C. Zhai and S. Wibowo, "A systematic review on artificial intelligence dialogue systems for enhancing English as foreign language students' interactional competence in the university," *Computers and Education: Artificial Intelligence*, vol. 4, p. 100134, 2023.

[6] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, et al., "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, pp. 102274, 2023.

[7] L. Wang, et al., "A survey on large language model based autonomous agents," *Frontiers of Computer Science*, vol. 18, no. 6, pp. 1-26, 2024.

[8] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *arXiv preprint arXiv:1804.04843*, 2018.

[9] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Medicine*, vol. 29, no. 8, pp. 1930-1940, 2023.

[10] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, et al., "Bloomberggpt: A large language model for finance," *arXiv preprint arXiv:2303.17564*, 2023.

[11] I. V. Serban, R. Lowe, P. Henderson, L. Charlin, and J. Pineau, "A survey of available corpora for building data-driven dialogue systems," *arXiv preprint arXiv:1512.05742*, 2015.

[12] M. Zaib, W. E. Zhang, Q. Z. Sheng, A. Mahmood, and Y. Zhang, "Conversational question answering: A survey," *Knowledge and Information Systems*, vol. 64, no. 12, pp. 3151-3195, 2022.

[13] C. Qu, L. Yang, C. Chen, M. Qiu, W. B. Croft, and M. Iyyer, "Open-retrieval conversational question answering," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 539-548.

[14] S. Reddy, D. Chen, and C. D. Manning, "Coqa: A conversational question answering challenge," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 249-266, 2019.

[15] Z. Dai, A. T. Chaganty, V. Y. Zhao, A. Amini, Q. M. Rashid, M. Green, and K. Guu, "Dialog inpainting: Turning documents into dialogs," in *International Conference on Machine Learning*, June 2022, pp. 4558-4586, PMLR.

[16] Y. Hwang, Y. Kim, H. Bae, J. Bang, H. Lee, and K. Jung, "Dialogizer: Context-aware Conversational-QA Dataset Generation from Textual Sources," *arXiv preprint arXiv:2311.07589*, 2023.

[17] A. Mohammadshahi, T. Scialom, M. Yazdani, P. Yanki, A. Fan, J. Henderson, and M. Saeidi, "Rquge: Reference-free metric for evaluating question generation by answering the question," *arXiv preprint arXiv:2211.01482*, 2022.

[18] X. Wang, B. Liu, S. Tang, and L. Wu, "QRelScore: Better Evaluating Generated Questions with Deeper Understanding of Context-aware Relevance," *arXiv preprint arXiv:2204.13921*, 2022.

[19] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "Gpteval: Nlg evaluation using gpt-4 with better human alignment," *arXiv preprint arXiv:2303.16634*, 2023.

[20] J. Wang, Y. Liang, F. Meng, Z. Sun, H. Shi, Z. Li, et al., "Is chatgpt a good nlg evaluator? a preliminary study," *arXiv preprint arXiv:2303.04048*, 2023.

[21] C. Qu, L. Yang, M. Qiu, Y. Zhang, C. Chen, W. B. Croft, and M. Iyyer, "Attentive history selection for conversational question answering," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, November 2019, pp. 1391-1400.

[22] M. Kaiser, R. Saha Roy, and G. Weikum, "Reinforcement learning from reformulations in conversational question answering over knowledge graphs," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 2021, pp. 459-469.

[23] Y. Li, W. Li, and L. Nie, "Mmcoqa: Conversational question answering over text, tables, and images," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, May 2022, pp. 4220-4231.

[24] N. S. K. Adatrao, G. R. Gadireddy, and J. Noh, "A

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3409377

survey on conversational search and applications in biomedicine," in *Proceedings of the 2023 ACM Southeast Conference*, April 2023, pp. 78-88.

[25] A. Saha, V. Pahuja, M. Khapra, K. Sankaranarayanan, and S. Chandar, "Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, April 2018.

[26] P. Christmann, R. Saha Roy, A. Abujabal, J. Singh, and G. Weikum, "Look before you hop: Conversational question answering over knowledge graphs using judicious context expansion," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, November 2019, pp. 729-738.

[27] Z. Chan, L. Liu, J. Li, H. Zhang, D. Zhao, S. Shi, and R. Yan, "Enhancing the open-domain dialogue evaluation in latent space," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, August 2021, pp. 4889-4900.

[28] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, July 2002, pp. 311-318.

[29] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, July 2004, pp. 74–81.

[30] S. Mehri and M. Eskenazi, "USR: An unsupervised and reference free evaluation metric for dialog generation," *arXiv preprint arXiv:2005.00456*, 2020.

[31] C. Tao, L. Mou, D. Zhao, and R. Yan, "Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 32, no. 1, April 2018.

[32] P. Achananuparp, X. Hu, and X. Shen, "The evaluation of sentence similarity measures," in *Data Warehousing and Knowledge Discovery: 10th International Conference, DaWaK 2008*, Turin, Italy, September 2-5, 2008, Springer Berlin Heidelberg, pp. 305–316.

[33] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, et al., "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877-1901, 2020.

[34] A. Stein and E. Maier, "Structuring collaborative information-seeking dialogues," *Knowledge-Based Systems*, vol. 8, no. 2-3, pp. 82–93, 1995.

[35] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1-67, 2020.

[36] B. Byrne, K. Krishnamoorthi, C. Sankar, A. Neelakantan, D. Duckworth, S. Yavuz, et al., "Taskmaster-1: Toward a realistic and diverse dialog dataset," *arXiv preprint arXiv:1909.05358*, 2019.

[37] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," *arXiv preprint arXiv:1710.03957*, 2017.

[38] C. Qu, L. Yang, C. Chen, M. Qiu, W. B. Croft, and M. Iyyer, "Open-retrieval conversational question answering," in *Proc. of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 2020, pp. 539–548.

[39] R. Anantha, S. Vakulenko, Z. Tu, S. Longpre, S. Pulman, and S. Chappidi, "Open-domain question answering goes conversational via question rewriting," *arXiv preprint arXiv:2010.04898*, 2020.

[40] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with BERT," *arXiv preprint arXiv:1904.09675*, 2019.

[41] Z. Mao and T. Nakagawa, "LEALLA: Learning lightweight language-agnostic sentence embeddings with knowledge distillation," *arXiv preprint arXiv:2302.08387*, 2023.

[42] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, and M. Zhang, "Towards general text embeddings with multi-stage contrastive learning," *arXiv preprint arXiv:2308.03281*, 2023.

[43] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[44] B. Pang, E. Nijkamp, W. Han, L. Zhou, Y. Liu, and K. Tu, "Towards holistic and automatic evaluation of open-domain dialogue generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, July 2020, pp. 3619-3629.

[45] L. Zheng, W. L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, et al., "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," in *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[46] H. Liang and H. Li, "Towards standard criteria for human evaluation of chatbots: A survey," *arXiv preprint arXiv:2105.11197*, 2021.

YONGIL KIM received the B.S. degree in electrical and computer engineering from Seoul University, South Korea, in 2019. He is currently pursuing the Ph.D. degree in Electrical and Computer Engineering with Seoul National University, South Korea. His research interest includes natural language processing.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3409377

YERIN HWANG received the B.S. degree in electrical and computer engineering from Seoul University, South Korea, in 2021. She is currently pursuing the Ph.D. degree in Interdisciplinary Program in Artificial Intelligence with Seoul National University, South Korea. Her research interest includes natural language processing.

HYUNKYUNG BAE earned her B.S. degree in Chemical Engineering from Pohang University of Science and Technology in 2015 and her M.S. degree in Electrical and Computer Engineering from Seoul University in 2022. She is currently working at LG AI Research in South Korea, focusing her research on natural language processing, particularly in language modeling.

TAEGWAN KANG received the Ph.D. degree in electrical and computer engineering from Seoul University, South Korea, in 2022. He currently works at LG AI Research, South Korea. His research interest includes natural language processing.

KYOMIN JUNG received the graduate degree from the Seoul Science High School, in February 1996, the B.Sc. degree from Seoul National University, in August 2003, and the Ph.D. degree from MIT, in June 2009. He worked at KAIST Computer Science Department, from 2009 to 2013, where he had joint appointments at the Department of Electrical Engineering and the Department of Mathematics. He is a Professor with the Electrical and Computer Engineering Department and an Adjunct Professor with the Department of Mathematical Sciences, Seoul National University. Currently, he is the Vice-Chair of the ECE Department for student affairs. His research interests include natural language processing, deep learning and applications, data analysis, and web services.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4