DialFill: Utilizing Dialogue Filling to Integrate Retrieved Knowledge in Responses

Xue, Qiang
Takiguchi, Tetsuya
Ariki, Yasuo

(Citation)
IEEE Access,13:61123-61135

(Issue Date)
2025-03-28

(Resource Type)
journal article

(Version)
Version of Record

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 March 2025, accepted 16 March 2025, date of publication 28 March 2025, date of current version 11 April 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3555650

RESEARCH ARTICLE

DialFill: Utilizing Dialogue Filling to Integrate Retrieved Knowledge in Responses

QIANG XUE, TETSUYA TAKIGUCHI, (Member, IEEE), AND YASUO ARIKI, (Life Member, IEEE)
Graduate School of System Informatics, Kobe University, Kobe 657-8501, Japan
Corresponding author: Qiang Xue (xueqiang@stu.kobe-u.ac.jp)

This work was supported by the Adaptable and Seamless Technology Transfer Program through Target-Driven Research and Development (A-STEP) from Japan Science and Technology Agency (JST) under Grant JPMJTR24RG.

ABSTRACT In knowledge-based dialogue systems, generating responses that are both contextually relevant and factually accurate requires efficient and precise integration of external knowledge. Pre-trained language models (LM-based) leverage extensive general knowledge but often struggle with accuracy in domain-specific or time-sensitive contexts due to their reliance on implicit knowledge. Conversely, knowledge-based approaches (KB-based) retrieve relevant information from external sources before response generation, yet they frequently fail to incorporate the retrieved content effectively, leading to responses that may omit critical information. To address these limitations, we propose **DialFill**, a novel response generation framework that reframes dialogue generation as a **Dialogue Filling** task. DialFill constructs an intermediate masked response that explicitly integrates the retrieved knowledge, subsequently predicting the missing components to ensure the final response incorporates all relevant information seamlessly. We validate DialFill on both unstructured (Wizard-of-Wikipedia) and structured (OpenDialKG) knowledge benchmarks, demonstrating competitive performance against state-of-the-art methods. In large language model experiments, DialFill significantly reduces the rate of retrieved content that is ignored, decreasing the number of ignored knowledge instances from 17.8% to 0.2%. These results highlight DialFill's potential to enhance the accuracy, reliability, and adaptability of knowledge-based dialogue systems, marking a significant advancement in the field.

INDEX TERMS Knowledge-based dialogue systems, external knowledge integration, dialogue filling.

I. INTRODUCTION

Building open-domain dialogue systems that generate human-like, factually accurate, and contextually relevant responses remains a central challenge in natural language processing [1], [2]. Recent advancements in large language models (LLMs) such as OpenAI's GPT-4 [3] and Meta's Llama3 [4] have demonstrated remarkable capabilities in generating coherent and fluent dialogue. These models encode extensive general knowledge within their parameters, enabling them to respond effectively without external retrieval mechanisms. However, they frequently produce hallucinations—responses that appear plausible but are factually incorrect or irrelevant—due to the static and limited scope of their training data [5], [6]. This limitation becomes particularly pronounced in domains requiring up-to-date or specialized knowledge, where their responses often lack necessary accuracy or relevance.

Knowledge-based (KB) dialogue systems have emerged as a promising approach to address these issues. By retrieving and integrating external knowledge from sources such as structured knowledge graphs (e.g. OpenDialKG [7]) or unstructured corpora (e.g. Wikipedia [8]), KB-based systems aim to generate more informative and grounded responses [8], [9]. Despite their strengths, these systems face challenges in seamlessly incorporating retrieved knowledge into dialogue responses. Often, KB-based methods overlook critical retrieved details or fail to align them effectively

The associate editor coordinating the review of this manuscript and approving it for publication was Fu Lee Wang.

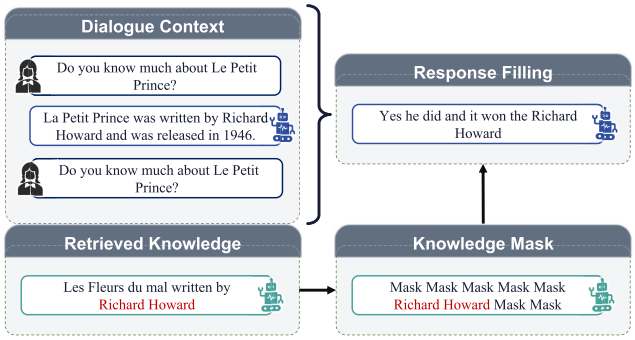


FIGURE 1. The overall flow of DialFill system.

with dialogue context, resulting in suboptimal coherence and informativeness [5], [10].

To address these challenges, we propose **DialFill**, a Dialogue Filling (DF-based) framework that extends traditional KB-based dialogue systems. DialFill reframes response generation into a three-step process: (i) extracting relevant keywords from retrieved knowledge (e.g. ''Richard Howard''); (ii) generating a masked response where only the extracted keyword remains unmasked (e.g. ''mask mask mask Richard Howard mask mask mask''); and (iii) completing the masked response to produce a coherent, knowledge-integrated output (e.g. ''Yes, he did and it won the Richard Howard''). This process, as illustrated in Figure 1, ensures the effective incorporation of external knowledge while seamlessly aligning it with the dialogue context. As an extension of KB-based methods, DialFill builds upon their strengths while addressing limitations in knowledge utilization by integrating retrieved information in a structured and context-aware manner.

We evaluate DialFill on two benchmark datasets: Wizard-of-Wikipedia, which requires unstructured knowledge integration, and OpenDialKG, which involves structured knowledge graph grounding [7], [8]. DialFill achieves state-of-the-art results across various metrics, outperforming both LM-based and KB-based baselines. Furthermore, our framework reduces the rate of ignored retrieved knowledge, demonstrating its effectiveness in leveraging external information for dialogue generation.

Our contributions are summarized as follows:

(i) We propose **DialFill**, a Dialogue Filling framework that extends KB-based dialogue systems by introducing masked response generation and completion, facilitating seamless integration of retrieved knowledge into responses.
(ii) We propose a unified training approach that combines *Keyword Prediction*, *Masked Response Generation*, and *Dialogue Filling* tasks, enhancing the alignment of retrieved knowledge with dialogue context.
(iii) We conduct extensive experiments on unstructured (Wizard-of-Wikipedia) and structured (OpenDialKG) datasets, demonstrating that DialFill outperforms conventional LM-based and KB-based methods, achieving state-of-the-art performance in response quality and knowledge integration.
(iv) We introduce an optimized inference mechanism, including keyword prediction and mask search strategies, that ensures robust and scalable performance across diverse knowledge formats and dialogue settings.

II. RELATED WORK

A. LANGUAGE MODEL-BASED DIALOGUE SYSTEMS

Recent advancements in large language models (LMs) have revolutionized generation-based dialogue systems, showcasing impressive capabilities in producing human-like responses. Notable models like OpenAI's GPT-4 [3] and Facebook's Llama3 [4] represent the latest generation of transformer-based models pre-trained on extensive datasets, allowing them to generate fluent and contextually relevant dialogue without relying on explicit retrieval mechanisms. Similarly, Meta's BlenderBot [11] and Google's LaMDA [12] further refined conversational quality by incorporating specialized pre-training and fine-tuning techniques that enhance user engagement and response coherence in open-domain dialogues.

Despite these advancements, a key challenge remains: LMs often produce responses with factual inaccuracies, a phenomenon known as ''hallucination'' [5], [6]. To address this, MixCL [13] uses contrastive learning to explicitly optimize the implicit knowledge elicitation process of LMs, effectively reducing hallucination in conversations. While LMs store extensive implicit knowledge, they also struggle with up-to-date or domain-specific knowledge integration due to static training data [14]. This limitation has prompted research into strategies such as retrieval-based augmentation [15], which reinforces factual accuracy by integrating external information, and exploring contrastive learning approaches to improve generation robustness [16].

B. KNOWLEDGE-BASED DIALOGUE SYSTEMS

Knowledge-based (KB) dialogue systems aim to address the limitations of purely generation-based approaches by integrating external knowledge sources to enhance coherence and factual grounding in responses [8], [9]. KB-based systems typically retrieve knowledge from structured sources like knowledge graphs (e.g. Freebase, Wikidata) and unstructured text corpora (e.g. Wikipedia), using information retrieval (IR) modules to identify contextually relevant information [5], [14]. Knowledge selection remains a primary focus in KB research, with recent works proposing advanced retrieval techniques such as GATE [17], a state-of-the-art knowledge selection method. GATE organizes knowledge selection into distinct phases—coupled with, after, and before generation—highlighting the advantages of selecting knowledge in advance to ease the burden on downstream response generation models, particularly large language models (LLMs).

While KB-based systems excel at grounding responses in factual content, they also face challenges, including retrieval

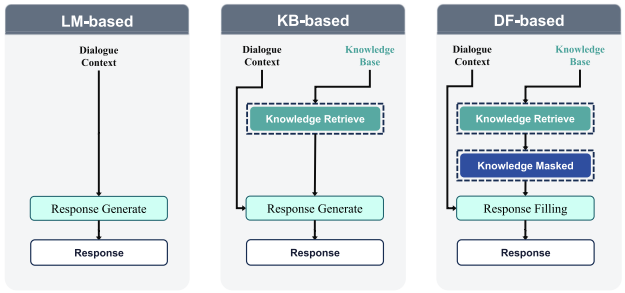


FIGURE 2. Comparison of three types of dialogue systems. (a) LM-based dialogue systems directly generate response. b) KB-based dialogue systems retrieves and directly incorporates knowledge into response generation. c) DF-based dialogue systems retrieves knowledge and fills masked responses for knowledge integration.

errors, inefficiency, and the integration of knowledge at multiple granularities [10], [18], [19]. Errors in retrieved content or the inability to seamlessly incorporate multiple sources of knowledge can result in responses that are verbose or lack critical information. Knowledge pre-selection methods, such as those explored in [20] and [21], address this issue by preparing relevant knowledge before response generation, enhancing response coherence and factuality. Our proposed framework, **DialFill**, extends this direction by rethinking response generation as a dynamic infilling task.

C. TEXT INFILLING

Text infilling is a natural language generation technique that involves predicting missing portions of text within a sentence or paragraph. This approach has been widely used in data augmentation, text summarization, and text editing tasks [22]. Pre-trained language models like BERT [23] and T5 [24] have demonstrated strong performance in text infilling scenarios, achieving contextual coherence in sentence completion and gap-filling tasks [25]. However, these models can sometimes generate generic or repetitive content and are sensitive to the positioning of masked tokens, which may limit coherence in complex dialogue contexts [22].

In our work, we adapt text infilling to the novel task of Dialogue Filling, where the goal is not simply to fill gaps in text but to dynamically generate responses that incorporate external knowledge relevant to the context. Unlike traditional text infilling, Dialogue Filling selects and integrates knowledge pre-retrieved from structured or unstructured sources, improving both the coherence and informativeness of responses. DialFill achieves this by coupling keyword prediction, masked response generation, and response completion in a unified multi-task framework, allowing for efficient and knowledge-base dialogue generation.

III. PROBLEM FORMULATION

Let x denote the dialogue context, y the target response, and k the ground-truth knowledge used during training. As illustrated in Fig. 2, given a knowledge corpus $\mathcal{K}$, the dialogue agent aims to predict an informative response y based on the context x, effectively incorporating relevant knowledge from $\mathcal{K}$. During the inference stage, the agent retrieves knowledge r from $\mathcal{K}$ to assist in generating the response.

A. LM-BASED METHODS

In the language-based (LM-based) dialogue systems, the model is trained on a dataset that includes knowledge annotations during the training phase. However, during the inference stage, as depicted in Fig. 2 (a), no external knowledge k is provided as input. Instead, the trained model generates responses based only on the dialogue context x.

B. KB-BASED METHODS

Traditional knowledge-based (KB-based) dialogue systems [8] follow a two-step process. As depicted in Fig. 2 (b), relevant knowledge is first retrieved from the knowledge corpus $\mathcal{K}$. In the second step, this retrieved knowledge is incorporated directly into the response generation process, with the response generator using this information to produce a coherent and contextually relevant response.

C. DF-BASED METHODS

We introduce the Dialogue Filling based (DF-based) dialogue systems as an extension of the KB-based. As illustrated in Fig. 2 (c), the relevant knowledge is first retrieved from $\mathcal{K}$. Next, a masked response is generated, where only the retrieved knowledge (e.g. 'Harry Potter') remains unmasked, while the rest of the sentence is fully masked (e.g. 'mask mask mask Harry Potter mask mask mask'). The model then completes the masked response by predicting the missing components, creating a coherent, knowledge-integrated output.

IV. PRELIMINARIES

We propose a DF-based dialogue agent for open-domain knowledge-based dialogue. The model, denoted $p_\theta(y|x)$, is based on a transformer-based language model architecture. It is initially trained on dialogue data, allowing it to generate informative responses given the input context x and the knowledge $\mathcal{K}$ during inference.

A. TRAINING STAGE

In the training stage, the model generates a response y given context x, following an explicit knowledge retrieval step [18], [26]. A maximum likelihood estimation (MLE) loss is employed to train the model using paired (x, y) dialogue data in a teacher-forcing setup [18], [26]:

$$\mathcal{L}_{\text{LM}} = -\log p_\theta(y|x) = -\sum_{t=1}^{|y|} \log p_\theta(y_t|y_{<t}, x). \quad (1)$$

In addition, our proposed model employs a multi-task learning approach, combining this MLE loss with other objectives to improve response quality.

B. INFERENCE STAGE

During inference, the model generates responses by feeding in context x and retrieved knowledge r, employing a greedy decoding strategy to predict the most probable tokens until an end-of-sequence (eos) token is reached.

V. DIALFILL

In this section, we present DialFill's training and inference stages. We outline each subcomponent and provide the relevant mathematical formulations to facilitate understanding of the model's learning objectives and operational framework.

A. TRAINING STAGE

The training phase of DialFill includes multiple tasks designed to predict appropriate keywords, generate masked responses, and ultimately fill in masked dialogue responses.

1) TASK 1: KEYWORD PREDICTION TASK

The goal of this task is to generate keywords relevant to the response context. In our method, keywords are entities within the response. As illustrated in Figure 3-Task 1, We use a *Named Entity Recognition* (NER) model[1] to extract entities within the target response, and randomly select one as the keyword kw. For responses lacking entities, we select a random token sequence as kw. This task employs an MLE loss over dialogue context x and ground-truth knowledge k:

$$\mathcal{L}_{\mathrm{KW}} = -\sum_{t=1}^{|kw|} \log p_\theta(kw_t | kw_{<t}, k, x). \tag{2}$$

2) TASK 2: MASKED RESPONSE PREDICTION TASK

This task aims to create a masked response of appropriate length based on kw. As illustrated in Figure 3-Task 2, the selected keyword kw remains unmasked, while the rest of the corresponding response y is masked. The masked response mr is then used as a target for training, with an MLE loss applied using inputs x, ground-truth knowledge k and keyword kw:

$$\mathcal{L}_{\mathrm{MR}} = -\sum_{t=1}^{|mr|} \log p_\theta(mr_t | mr_{<t}, k, x, kw). \tag{3}$$

3) TASK 3: DIALOGUE FILLING TASK

The final task predicts the masked portions of the response to generate a complete, knowledge-integrated response. As shown in Figure 3-Task 3, the model is trained using the ground-truth response y as a label, with k, x, kw, and mr as inputs. The MLE loss for this task is:

$$\mathcal{L}_{\mathrm{DF}} = -\sum_{t=1}^{|y|} \log p_\theta(y_t | y_{<t}, k, x, kw, mr). \tag{4}$$

[1] https://spacy.io/api/entityrecognizer/

4) OPTIMIZATION

During training, the final training objective is defined as:

$$\mathcal{J}(\theta) = \alpha_1 \mathcal{L}_{\mathrm{LM}} + \alpha_2(\mathcal{L}_{\mathrm{KW}} + \mathcal{L}_{\mathrm{MR}} + \mathcal{L}_{\mathrm{DF}}), \tag{5}$$

where four losses are optimized jointly and α_1, α_2 denote the weights of the four losses, respectively.

B. INFERENCE STAGE

The inference stage in DialFill consists of three steps: *Keyword Generation*, *Masked Response Generation*, and *Dialogue Filling*.

1) STEP 1: KEYWORD GENERATION

In this step, the model identifies a keyword k relevant to the dialogue context x and retrieved knowledge r. Here, r refers to information retrieved by a knowledge retrieval model, which can be structured or unstructured, depending on the source and format of the data. The approach for keyword identification adapts accordingly.

- **Structured Knowledge:**
  For structured knowledge (e.g. knowledge triples), the system directly uses the target entity in the knowledge tuple as the keyword kw. For instance, given a tuple such as (Les Fleurs du mal, written by, Richard Howard), the keyword k is identified as "Richard Howard." This straightforward selection process bypasses additional keyword prediction methods.
- **Unstructured Knowledge:**
  For unstructured knowledge (e.g. free-form text), the model attempts to detect named entities using NER model.
  - If named entities are detected, the system selects one entity at random as the keyword kw.
  - If no entities are detected, the model relies on *keyword prediction* based on the dialogue context x and retrieved knowledge r.

  In *Keyword Prediction* (Figure 4-P), the model generates a keyword kw by analyzing the dialogue context x and retrieved knowledge r, following Equation 2 from the training stage, which was used to predict keywords by maximizing their relevance in the context of the response.

2) STEP 2: MASKED RESPONSE GENERATION

After determining the keyword kw, the model generates a masked response mr that surrounds kw with a sequence of mask tokens, setting up the context for dialogue filling.

- **Mask Append** (Figure 4-A):
  If the model has not been explicitly trained to generate masked responses (i.e., it has not been trained using Equation 3 for masked response prediction), it cannot directly predict where to place the mask tokens accurately. In such cases, the system appends a random number of mask tokens on each side of the keyword kw, constrained by the hyperparameters L_{min} and L_{max}. This approach approximates the masked response m, albeit without refined placement.

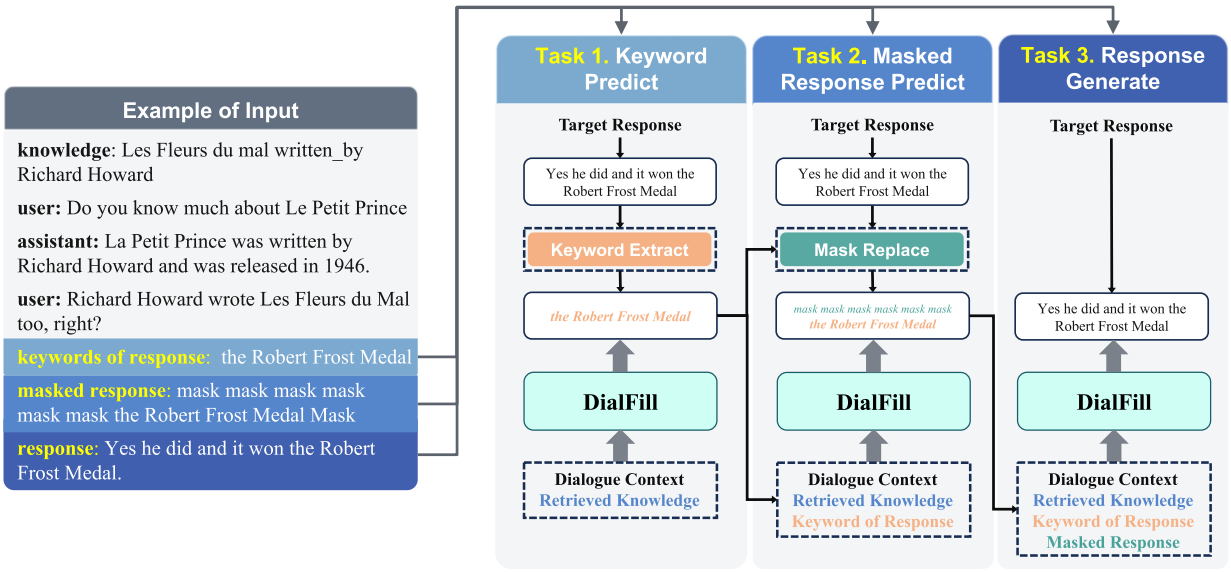


FIGURE 3. Overview of the training stage of DialFill. DialFill consists of three tasks: (i) Keyword Prediction Task, where the model(DialFill) extracts key information relevant to the response; (ii) Masked Response Prediction, where the model(DialFill) generates a partially masked version of the response based on the extracted keyword; and (iii) Dialogue Filling, where the model(DialFill) completes the masked response to produce the final coherent response.

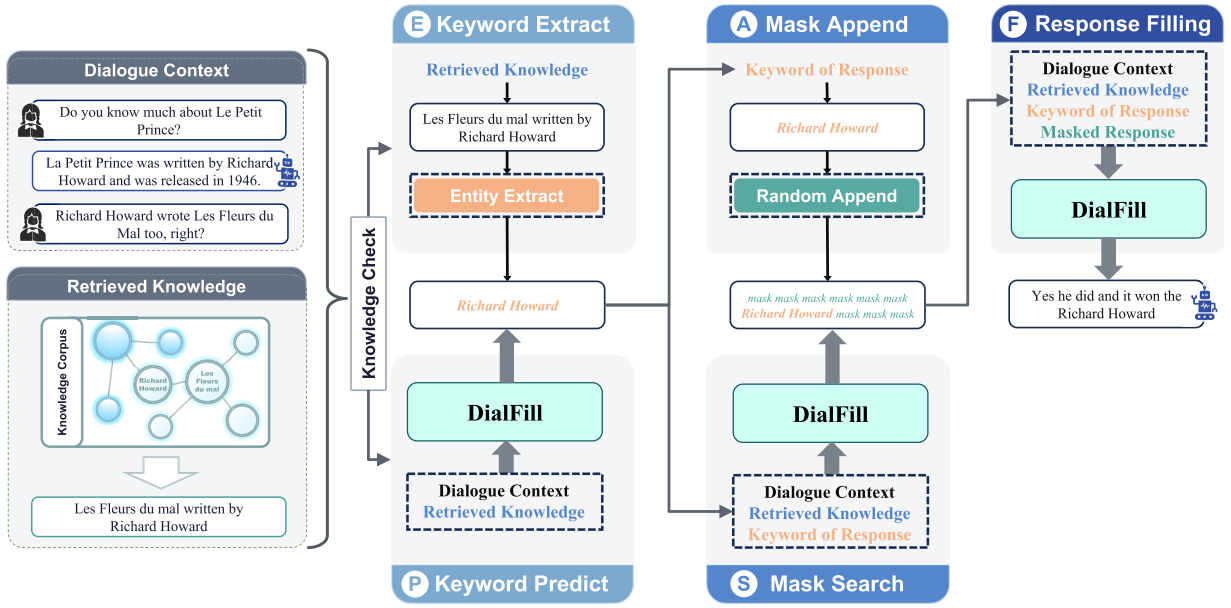


FIGURE 4. Overview of the inference stage of DialFill. The stage consists of three steps: (i) Keyword Generation, where relevant keywords are identified from the retrieved knowledge (E) or predicted based on the dialogue context (P); (ii) Masked Response Generation, which involves placing mask tokens around the keyword to prepare for response generation, using either a random append method (A) or an optimized mask search strategy (S); and (iii) Dialogue Filling, where the model completes the masked response to produce a coherent, knowledge-integrated response.

- **Mask Search** (Figure 4-S):
  For models trained with masked response generation, the *Mask Search Strategy* (see Appendix A) is used to determine the optimal number of mask tokens on each side of *kw*. The Mask Search Strategy evaluates the token generation probabilities to adjust mask length dynamically, maximizing coherence in the resulting response.
  Specifically, the algorithm iteratively increases the length of the left and right masks, using the model's probability predictions to identify the mask lengths that yield the highest likelihood of a coherent response. This method optimizes both left and right mask lengths (L_{left} and L_{right}) to produce a well-formed masked response *mr*.

3) STEP 3: DIALOGUE FILLING (FIGURE 4-f)

In the final step, the model fills in the masked portions of *mr* based on *x*, *r*, *kw*, generating a coherent and knowledge-integrated response *y*. This process follows the

setup from Equation 4, allowing the model to produce a response y that aligns with the context and incorporates relevant knowledge.

VI. EXPERIMENTAL SETUP

A. DATASETS

We evaluate our proposed DialFill on two widely used knowledge-base dialogue datasets: the unstructured knowledge dataset **Wizard of Wikipedia** (WoW) [8] and the structured knowledge dataset **OpenDialKG** [7]. These datasets encompass diverse conversational scenarios and knowledge formats, providing a comprehensive basis for assessing the effectiveness of our approach.

1) WIZARD OF WIKIPEDIA (WoW)

is a dialogue dataset constructed through crowd-sourcing, where Wikipedia serves as the primary knowledge source. In WoW, two participants engage as a *wizard* and an *apprentice*. The wizard responds by selecting appropriate knowledge sentences from Wikipedia to inform the conversation. The test set of WoW is divided into *test seen* and *test unseen* categories, depending on whether the topic appears in the training set.

2) OpenDialKG

is a dataset comprising open-domain dialogues grounded in a knowledge graph. Each dialogue turn is associated with annotated reasoning paths, enabling the use of structured graph information during conversation. Since OpenDialKG does not provide an official test split, following previous work [17], we partition it into *seen* and *unseen* categories, consistent with the methodology used for WoW.

B. EVALUATION METRICS

To comprehensively evaluate our approach, we employ a suite of metrics that assess both the linguistic quality of the generated responses and their relevance to the underlying knowledge. These metrics are categorized into three groups: *Response Quality Metrics*, *Target Knowledge Relevance Metrics*, and *Retrieved Knowledge Relevance Metrics*.

1) RESPONSE QUALITY METRICS

Evaluate the linguistic quality and fluency of the generated responses by comparing them to the ground truth.

- **F1** [8]: Calculates the unigram F1 score between the generated responses and the ground truth, measuring lexical overlap.
- **ROUGE-L** (RL) [27]: Measures the longest common subsequence between the generated and reference responses, capturing structural similarity.
- **BLEU-4** (B4) [28]: Evaluates the 4-gram precision of the generated responses against the ground truth.
- **METEOR** (MT) [29]: Computes a score based on unigram matches between the generated responses and the reference, considering synonyms and stemming.

2) TARGET KNOWLEDGE RELEVANCE METRICS

Assess how well the generated responses align with the ground-truth knowledge associated with each dialogue turn.

- **Knowledge-F1** (KF1) [8]: Calculates the F1 score between the generated responses and the ground-truth knowledge sentences, indicating the informativeness of the responses.
- **Entity-F1** (EF1): Identifies named entities in both the generated responses and the ground truth using spaCy, and computes the F1 score based on matched entities to evaluate accurate entity usage.

3) RETRIEVED KNOWLEDGE RELEVANCE METRICS

Evaluate the extent to which the retrieved knowledge is incorporated into the generated responses.

- **Retrieved-F1** (RF1) [proposed]: Computes the F1 score between the knowledge retrieved and the responses generated, reflecting the integration of retrieved content.
- **Zero-Retrieved** (ZR) [proposed]: Calculates the proportion of generated responses that have an F1 score of zero with the retrieved knowledge, indicating the absence of retrieved knowledge in the responses.
  - For WoW: Measures the F1 overlap between the retrieved knowledge sentences and the generated responses.
  - For OpenDialKG: Since the knowledge is represented as triples (subject, relation, object), we consider only the *object* as the target knowledge for the F1 calculation, as it is the primary piece of information the dialogue system should incorporate.

C. BASELINES

We compare DialFill with a variety of baseline methods, each using their *own* generation strategy and architecture. For instance, GATE retrieves knowledge but uses a standard GPT-2-small fine-tuned on the dataset for generation, whereas KnowledGPT uses a BERT-based retriever plus GPT-2 with a reinforcement objective. In contrast, DialFill fine-tunes its own GPT-2 or Llama3 model under a multi-task paradigm that includes masked response generation. These baselines are categorized into *Language Model (LM)-based Methods* and *Knowledge Base (KB)-based Methods.*

1) LM-BASED METHODS

- **GPT-2** [26]: Fine-tunes GPT-2-small on dialogue data without utilizing external knowledge.
- **BlenderBot** [11]: Pre-trains a Transformer with an encoder-decoder architecture on Reddit data and fine-tunes it on knowledge-base dialogue data.
- **KnowExpert** [18]: Adapts GPT-2-small for open-domain dialogues using a topic-aware adapter that groups Wikipedia articles via topic modeling and employs a mixture-of-adapters architecture.
- **MSDP** [10]: Utilizes a multi-stage prompting approach, designing task-specific prompts with instructions and in-context examples, and leverages Megatron-LM [30] to generate knowledge and responses in a two-stage process.

- **MixCL** [13]: Introduces a mixed contrastive learning objective to optimize the implicit knowledge elicitation process in LMs, effectively reducing hallucination and improving factuality in knowledge-base dialogues.

2) KB-BASED METHODS

- **DiffKG** [31]: it employs Transformer to generate relation sequences on KG and generates responses based on retrieved entities.
- **NPH** [32]: Reduces hallucinations by retrieving relevant entities through propagation of a crafted query signal over a knowledge graph to refine knowledge selection.
- **TMN** [8]: Combines a Transformer with an external memory network to select knowledge and generate responses.
- **DukeNet** [33]: Employs a dual learning scheme to model both knowledge shift and response generation without relying on pre-trained language models.
- **KnowledGPT** [20]: Leverages pre-trained language models in a KB-based approach, using BERT for knowledge selection and GPT-2 for response generation, optimized jointly with reinforcement learning.
- **KnowBART**: Selects knowledge using RoBERTa and generates responses using BART-Large.
- **GATE** [17]: Introduces a generator-agnostic knowledge selection method that identifies contextually relevant knowledge prior to response generation, enhancing the informativeness of responses while alleviating the burden on subsequent generation models. It utilizes GPT-2-small as the response generator.

D. PROPOSED METHODS

To evaluate the effectiveness of DialFill and analyze the impact of its components, we design several variants:

1) BASE MODELS

- **DialFill$^{\mathbf{L}}$**: A version of DialFill that operates without external knowledge (**L** indicates *Language Model*). It is trained solely on dialogue context using the multi-task objective in Equation 5, relying on the model's parameters during both training and inference.
- **DialFill$^{\mathbf{K}}$**: A version of DialFill that incorporates external knowledge (**K** indicates *Knowledge*). During training, it integrates the target knowledge into the multi-task objective in Equation 5. During inference, it uses GATE [17], a state-of-the-art knowledge retrieval model, to retrieve relevant knowledge, which is combined with the dialogue context for response generation.

2) DF-BASED METHODS

To further investigate the impact of the multi-task learning design, we introduce three variants that utilize different inference strategies:

- **DialFill$^{\mathbf{L}}$-PSF**: A variant of DialFill$^{\mathrm{L}}$ that frames response generation as a three-step process: (i) predicting keywords (Figure 2-P), (ii) generating mask tokens around these keywords (Figure 2-S), and (iii) filling the masked response to produce the final output (Figure 2-F). Since this model does not use external knowledge, all steps rely on the dialogue context and learned model parameters.
- **DialFill$^{\mathbf{K}}$-PSF**: Similar to DialFill$^{\mathrm{L}}$-PSF but incorporates external knowledge during both training and inference. It follows the same three-step process, using the knowledge retrieved from GATE alongside the dialogue context.
- **DF-based methods ensure robustness and scalability**: DialFill$^{\mathrm{K}}$-ESF includes a fallback mechanism that defaults to DialFill$^{\mathrm{K}}$-PSF when no entities are detected or no relevant knowledge is retrieved. This design ensures robust performance even in cases of limited or unavailable external knowledge. As a result, DialFill$^{\mathrm{K}}$-ESF achieves comparable or slightly better performance than DialFill$^{\mathrm{K}}$-PSF across most metrics (e.g. F1 = 23.7 vs. 23.5 on WoW Test Seen), demonstrating its scalability and effectiveness across varied inference scenarios.

E. IMPLEMENTATION DETAILS

We implement our models using the HuggingFace Transformers library, using GPT-2-Small [26][2] and Llama3-8B-Instruct [4][3] as the base architectures. Although Llama3 generally achieves stronger results, we also employ GPT-2 for three main reasons: (i) It is more feasible to run on limited hardware, (ii) Many prior works on WoW and OpenDialKG rely on GPT-2-Small or models of similar size, ensuring fair comparisons, (iii) Larger models illustrate the upper bound of performance under more generous computational resources.

For GPT-2-Small, we fine-tune the model with a batch size of 4 and a learning rate of 6×10^{-5} on a single NVIDIA GeForce RTX 2070 GPU. For Llama3-8B-Instruct, we fine-tune the model using Low-Rank Adaptation (LoRA) [35], with a batch size of 4 and a learning rate of 4×10^{-4} on a single NVIDIA GeForce RTX A6000 GPU. Hyperparameters are selected based on preliminary experiments.

In Equation 5, we set the loss weights at $\alpha_1 = 0.4$ and $\alpha_2 = 0.6$. Model checkpoints are selected based on the performance of the validation set evaluated after each epoch. Following prior work [36], we set the minimum and maximum mask lengths, $L_{\min}$ and $L_{\max}$, to 5 and 10, respectively. We employ GATE as the knowledge retrieval model for both WoW and OpenDialKG datasets due to its superior performance in knowledge retrieval. During testing, we use greedy decoding for response generation.

VII. EXPERIMENTAL RESULTS

A. RESPONSE QUALITY EVALUATION

Table 1 compares LM-based, KB-based, and our DF-based methods on both the WoW and OpenDialKG datasets. Overall, DF-based methods consistently achieve higher response quality than their LM-based and KB-based counterparts. We note that, while knowledge-grounded approaches typically outperform language models that rely solely on

[2]https://huggingface.co/openai-community/gpt2
[3]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

TABLE 1. The evaluation results of response generation on WoW and OpenDialKG datasets. The baseline model results for WoW are reported from [34]. We highlight the results of DialFill that significantly exceed the previous-best methods in boldface (t-test, $p < 0.05$). We also highlight the best results of previous KB-based methods and LM-based methods by underlining them, respectively.

Type	Method	Test Seen						Test Unseen					
		F1	RL	B4	MT	KF1	EF1	F1	RL	B4	MT	KF1	EF1
WOW													
LM-based	GPT-2 [26]	16.5	17.3	0.7	14.6	11.8	5.9	15.5	16.5	0.4	13.6	10.4	4.3
	BlenderBot [11]	18.8	19.4	2.3	18.0	18.2	13.1	17.8	16.9	0.8	15.0	15.7	7.1
	KnowExpert [18]	18.7	18.6	1.3	16.5	14.1	9.8	16.7	17.2	0.6	14.5	11.8	5.5
	MSDP [10]	17.8	16.5	1.9	18.2	21.7	13.9	16.9	16.1	1.1	16.2	20.3	8.4
	MixCL [13]	21.6	20.5	2.7	20.5	22.3	16.3	19.6	18.8	1.4	18.0	18.0	11.6
KB-based	TMN [35]	17.3	17.0	1.1	14.8	15.8	8.7	14.4	14.5	0.3	11.5	9.4	2.1
	DukeNet [33]	18.5	17.7	1.9	16.0	18.5	12.0	15.9	15.9	1.1	13.7	14.7	8.0
	KnowledGPT [20]	21.1	20.1	3.4	20.0	22.2	15.5	19.5	18.4	2.6	18.3	20.0	11.7
	KnowBART	21.1	18.9	3.3	17.8	21.3	16.2	21.0	18.3	3.6	17.9	22.5	16.2
	GATE [17]	23.2	21.9	4.3	21.1	24.3	19.5	21.2	20.3	3.4	19.3	21.4	16.7
DF-based	**DialFill^{L}**-PSF	15.9	16.9	0.5	13.7	11.6	8.5	15.1	16.5	0.4	13.0	10.8	6.7
	DialFill^{K}-PSF	23.5	21.7	4.3	20.6	24.5	**22.1**	21.5	20.4	3.6	18.7	21.7	**19.2**
	DialFill^{K}-ESF	**23.7**	**22.2**	**4.6**	**21.7**	**26.7**	21.5	**22.0**	**21.2**	**4.0**	**20.3**	**23.9**	**19.2**
OpenDialKG													
LM-based	GPT-2 [26]	21.4	22.1	2.4	19.5	11.7	11.4	20.4	21.1	1.8	18.4	10.1	7.9
	MixCL [13]	21.6	22.3	2.5	19.8	12.0	9.7	20.7	21.5	2.0	19.3	10.7	8.1
KB-based	DiffKG [31]	17.5	18.0	1.2	14.8	14.2	8.1	17.9	17.8	1.2	14.8	14.1	7.2
	NPH [32]	24.1	23.3	3.2	21.4	14.8	15.9	22.5	21.9	2.5	19.8	12.4	11.4
	GATE [17]	25.1	23.9	4.2	22.1	20.5	23.5	24.3	23.2	3.7	21.6	19.2	21.1
DF-based	**DialFill^{L}**-PSF	21.8	22.7	2.4	20.4	14.9	13.3	20.8	22.0	2.1	19.7	13.5	10.2
	DialFill^{K}-PSF	26.6	**25.8**	4.8	23.9	**23.7**	22.8	25.1	24.5	**4.3**	22.8	**22.8**	20.7
	DialFill^{K}-ESF	**26.7**	**25.8**	**5.0**	**24.1**	22.3	22.5	**25.3**	**24.7**	4.2	**23.3**	21.1	19.4

parametric knowledge, certain LM-based methods (e.g., MixCL) are still able to surpass some KB-based baselines, indicating that a strong model design and training strategy can compensate for the lack of explicit retrieval. Nevertheless, by integrating external knowledge into a multi-step Dialogue Filling procedure, our approach achieves further gains. This synergy between retrieval and multi-task generation helps bridge knowledge gaps and yield more coherent, factually grounded responses. We make the following observations:

- **DF-based methods outperform KB-based methods**: DialFill^{K}-ESF achieves an F1 score of 23.7 on WoW Test Seen and 26.7 on OpenDialKG Test Seen, surpassing GATE, the best KB-based baseline, by 0.5 and 1.6 points, respectively. This demonstrates the effectiveness of Dialogue Filling in leveraging retrieved knowledge more efficiently.
- **Knowledge significantly improves DF-based models**: DialFill^{K}-PSF achieves an F1 score of 23.5 on WoW Test Seen, outperforming DialFill^{L}-PSF, which does not use external knowledge, by 7.6 points (F1 = 15.9). This large gap highlights the critical role of external knowledge in generating accurate and informative responses.
- **Enhanced inference boosts performance**: DialFill^{K}-ESF outperforms DialFill^{K}-PSF across most metrics, achieving an F1 score of 23.7 on WoW Test Seen and 22.0 on Test Unseen, compared to 23.5 and 21.5 for DialFill^{K}-PSF. The slight improvement validates the effectiveness of optimized inference strategies, such as mask generation and entity-based keyword extraction, in generating more coherent and contextually relevant responses.
- **DF-based methods ensure robustness and scalability**: The fallback mechanism in DialFill^{K}-ESF enables robust performance even when retrieved knowledge is limited or unavailable, achieving results comparable to DialFill^{K}-PSF while maintaining flexibility across varied conditions.

B. RETRIEVED KNOWLEDGE RELEVANCE EVALUATION

Table 2 evaluates retrieved knowledge relevance using RF1 and ZR metrics. DF-based methods outperform KB-based baselines in RF1, indicating superior integration of retrieved knowledge. For instance, DialFill^{K}-ESF achieves an RF1 of 51.8 on WoW Test Seen, surpassing GATE (45.9). On OpenDialKG, DialFill^{K}-ESF improves RF1 by 1.9 points compared to GATE, confirming its enhanced knowledge integration capabilities.

Interestingly, DialFill^{K}-PSF achieves slightly better ZR scores than DialFill^{K}-ESF (e.g. 5.8 vs. 6.2 on OpenDialKG Test Seen). This is likely because DialFill^{K}-PSF always incorporates some retrieved knowledge due to its simpler inference strategy. In contrast, DialFill^{K}-ESF's enhanced

TABLE 2. Evaluation results with RF1↑ (higher is better) and ZR↓ (lower is better) metrics to assess knowledge incorporation in generated responses.

Type	Method	Test seen		Test unseen	
		RF1↑	ZR↓	RF1↑	ZR↓
WOW					
KB-based	GATE	45.9	5.0	43.3	4.5
DF-based	**DialFill^{K}**-PSF	47.1	4.8	45.2	4.4
	DialFill^{K}-ESF	**51.8**	**4.7**	**50.7**	**3.9**
OpendialKG					
KB-based	GATE	21.1	11.4	20.7	11.3
DF-based	**DialFill^{K}**-PSF	21.2	**5.8**	20.7	**4.9**
	DialFill^{F}-ESF	**23.0**	6.2	**22.8**	6.3

mechanisms, such as fallback strategies, occasionally prioritize fluency or coherence, leading to slightly higher ZR but better overall RF1 and response quality. These results highlight that DialFill^{K}-ESF effectively balances knowledge integration and response coherence.

C. RESULTS IN LLM

Table 3 shows that our DF-based methods significantly outperform KB-based baselines in both zero-shot and fine-tuned settings. As before, each KB-based method relies on its original generator, while DF-based uses a multi-task generator. We highlight two variants: DialFill-PAF, which Predicts keywords from the context, and DialFill-EAF, which Extracts keywords directly from retrieved knowledge. In zero-shot mode, DialFill-PAF has not learned to identify the best knowledge-focused keywords and thus underperforms DialFill-EAF. After fine-tuning, both methods learn to align keywords with retrieved knowledge, reducing the performance gap significantly. We make the following observations:

- **DF-based methods achieve superior knowledge incorporation**: DialFill^{K}-ESF achieves the lowest ZR values, with 0.2 on Test Seen and 0.3 on Test Unseen in the fine-tuning setting, indicating nearly all responses integrate retrieved knowledge. These values are markedly lower than both GATE and DialFill^{K}-PSF, showing the effectiveness of the enhanced inference pipeline in our framework.
- **LLMs boost DF-based performance across metrics**: DialFill^{K}-ESF achieves the highest F1 (28.3 on Test Seen, 28.0 on Test Unseen) and RF1 (24.0 on Test Seen, 23.6 on Test Unseen), significantly outperforming GATE (F1 = 27.3, RF1 = 17.3 on Test Seen). These results confirm the synergy between LLMs and our DF-based framework, which maximally leverages the generative capacity of large models.
- **DF-based methods ensure robustness even in zero-shot settings**: DialFill^{K}-EAF outperforms all baselines in zero-shot conditions, achieving an F1 score of 18.7 on Test Seen compared to GATE's 12.3. This demonstrates the robustness of our framework in incorporating knowledge and generating accurate responses without task-specific training.

D. ANALYSIS ON KEYWORD PREDICTION AND MASKED RESPONSE GENERATION

To assess whether the proposed keyword prediction (KW) and masked response generation (MR) steps are effective, we evaluate two additional metrics:

KW_{UR} (Keyword Usage Rate): the percentage of final responses that include the predicted keyword. A high KW_{UR} indicates that the model successfully preserves and uses the predicted keyword in its final answer.

MR_{LR} (Masked Response Length Ratio): the average absolute difference in length between the generated masked response (Task 2) and the final filled response (Task 3). A smaller gap suggests stable expansions of the masked skeleton, indicating that the masked response serves as a coherent scaffold for the final output.

Table 4 shows KW_{UR} and MR_{LR} for DialFill. We see that 76–99% of predicted keywords appear in the final responses, confirming that the model effectively leverages them. Meanwhile, MR_{LR} values remain around 3–5 words difference, indicating that the model does not deviate drastically from the masked skeleton, thus maintaining coherent context around the keyword. Nevertheless, each sub-step (keyword prediction and masked response generation) can introduce errors that may propagate into the final response. This effect is more pronounced in smaller models like GPT-2, where KW_{UR} can drop to 76% and MR_{LR} can reach 5.40 words. In contrast, larger models such as Llama3 achieve KW_{UR} values as high as 99% and MR_{LR} as low as 2.74, indicating that they more faithfully adhere to the intermediate constraints. These findings suggest that while multi-step generation provides a structured way of integrating knowledge, larger models can better maintain consistency with the predicted keywords and masked response skeletons throughout the Dialogue Filling process.

E. ABLATION STUDIES

Table 5 presents the results of the ablation studies, comparing the base model with several variants to analyze the impact of different components. The findings are as follows:

No $\mathcal{L}_{LM}$ – Removing the general language modeling loss leads to noticeable performance drops, particularly in F1 (down by 0.9 points on OpenDialKG Test Seen). This indicates that $\mathcal{L}_{LM}$ contributes to learning generic language features that enhance response fluency and coherence.

No $\mathcal{L}_{KW}$ – Excluding the keyword prediction task results in minor declines (e.g. F1 decreases by 0.1 on OpenDialKG Test Seen). The relatively small impact suggests that keyword prediction plays a supplementary role in guiding the model to relevant content without being the primary driver of response quality.

No $\mathcal{L}_{MR}$ – Removing the masked response prediction task causes moderate drops, especially in KF1 (down by 0.9 on

TABLE 3. The evaluation results of response generation on OpenDialKG datasets with LLM. We highlight the results of DialFill that significantly exceed the previous-best methods in **boldface** (t-test, $p < 0.05$).

Type	Method	Test Seen								Test Unseen							
		F1	RL	B4	MT	KF1	EF1	RF1	ZR	F1	RL	B4	MT	KF1	EF1	RF1	ZR
Zero-Shot																	
KB-based	GATE	12.3	10.2	0.6	15.5	7.6	17.1	6.4	18.0	11.6	9.8	0.5	14.7	6.7	15.9	5.2	21.1
DF-based	**DialFill^{K}**-PAF	14.8	11.9	1.1	12.9	14.6	14.2	15.3	21.8	14.2	11.4	1.0	12.8	13.9	14.4	14.9	26.5
	DialFill^{K}-EAF	**18.7**	**15.9**	**1.9**	**18.8**	**18.9**	**24.1**	**20.9**	**9.6**	**18.2**	**15.4**	**1.6**	**18.4**	**18.7**	**20.7**	**20.7**	**9.5**
Fine-Tuning																	
KB-based	GATE	27.3	25.8	4.7	24.4	18.1	21.7	17.3	17.8	26.9	25.4	4.3	24.2	17.7	21.7	17.6	17.7
DF-based	**DialFill^{K}**-PSF	28.6	**27.3**	**5.7**	26.6	20.5	25.8	19.5	5.0	27.6	26.3	4.8	25.5	**19.2**	23.4	19.4	4.5
	DialFill^{K}-ESF	**28.3**	26.8	**5.7**	**28.1**	**21.0**	**27.1**	**24.0**	**0.2**	**28.0**	**26.4**	**5.1**	**27.7**	18.9	**24.7**	**23.6**	**0.3**

TABLE 4. Analysis of keyword prediction and masked response generation on OpenDialKG dataset. See section VII-D.

Method	Test seen		Test unseen	
	$KW_{UR} \uparrow$	$MR_{LR} \downarrow$	$KW_{UR} \uparrow$	$MR_{LR} \downarrow$
Fine-Tuned GPT-2				
DialFill^{K}-PSF	85.97%	5.37	86.34%	5.40
DialFill^{K}-ESF	76.42%	3.99	78.22%	3.87
Fine-Tuned Llama-3				
DialFill^{K}-PSF	99.33	5.45	99.47	5.53
DialFill^{F}-ESF	95.23	2.74	97.37	2.65

TABLE 5. Ablation study. The base model, DialFill, is compared with several variants. See Section VII-E.

Methods	Test seen			Test unseen		
	F1	B4	KF1	F1	B4	KF1
WOW						
Base model	23.7	4.6	26.7	22.0	4.0	23.9
-w/o $\mathcal{L}_{LM}$	$23.5_{\downarrow 0.2}$	$4.3_{\downarrow 0.3}$	$26.2_{\downarrow 0.5}$	$21.4_{\downarrow 0.6}$	$3.6_{\downarrow 0.4}$	$23.3_{\downarrow 0.6}$
-w/o $\mathcal{L}_{KW}$	$23.5_{\downarrow 0.2}$	$4.6_{\downarrow 0.0}$	$26.2_{\downarrow 0.5}$	$21.7_{\downarrow 0.3}$	$4.0_{\downarrow 0.0}$	$23.3_{\downarrow 0.6}$
-w/o $\mathcal{L}_{MR}$	$23.3_{\downarrow 0.4}$	$4.4_{\downarrow 0.2}$	$26.2_{\downarrow 0.5}$	$21.2_{\downarrow 0.8}$	$3.6_{\downarrow 0.4}$	$23.0_{\downarrow 0.9}$
-w/o $\mathcal{L}_{DF}$	$0.0_{\downarrow 23.7}$	$0.0_{\downarrow 4.6}$	$0.0_{\downarrow 26.7}$	$0.0_{\downarrow 22.0}$	$0.0_{\downarrow 4.0}$	$0.0_{\downarrow 23.9}$
-w/o $\mathcal{A}_{MS}$	$23.4_{\downarrow 0.3}$	$4.6_{\downarrow 0.0}$	$26.1_{\downarrow 0.7}$	$21.3_{\downarrow 0.7}$	$3.7_{\downarrow 0.3}$	$23.0_{\downarrow 0.9}$
OpenDialKG						
Base model	26.7	5.0	22.3	25.3	4.2	21.1
-w/o $\mathcal{L}_{LM}$	$25.8_{\downarrow 0.9}$	$4.6_{\downarrow 0.4}$	$20.6_{\downarrow 1.7}$	$25.1_{\downarrow 0.2}$	$4.1_{\downarrow 0.1}$	$18.7_{\downarrow 2.4}$
-w/o $\mathcal{L}_{KW}$	$26.6_{\downarrow 0.1}$	$4.7_{\downarrow 0.3}$	$21.8_{\downarrow 0.5}$	$25.3_{\downarrow 0.0}$	$3.7_{\downarrow 0.5}$	$20.6_{\downarrow 0.5}$
-w/o $\mathcal{L}_{MR}$	$26.2_{\downarrow 0.5}$	$4.4_{\downarrow 0.6}$	$21.3_{\downarrow 1.0}$	$25.3_{\downarrow 0.0}$	$4.1_{\downarrow 0.1}$	$20.0_{\downarrow 1.1}$
-w/o $\mathcal{L}_{DF}$	$0.0_{\downarrow 26.7}$	$0.0_{\downarrow 5.0}$	$0.0_{\downarrow 22.3}$	$0.0_{\downarrow 25.3}$	$0.0_{\downarrow 4.2}$	$0.0_{\downarrow 21.1}$
-w/o $\mathcal{A}_{MS}$	$25.8_{\downarrow 0.9}$	$4.5_{\downarrow 0.5}$	$23.4_{\uparrow 1.1}$	$24.8_{\downarrow 0.5}$	$4.1_{\downarrow 0.1}$	$21.9_{\uparrow 0.8}$

WoW Test Unseen and 1.1 on OpenDialKG Test Unseen). This demonstrates that masked response prediction helps the model structure responses around relevant knowledge, enhancing factual accuracy.

No $\mathcal{L}_{DF}$ – We observe that removing Dialogue Filling (DF) task completely collapses the generation (F1=0), since DF is responsible for synthesizing the final coherent response. By comparison, removing Keyword Prediction (KW) task or Masked Response Generation (MR) task causes a smaller but still noticeable degradation in both F1 and knowledge metrics. Hence, while DF is critical for ensuring the final text is well-formed, the KW and MR steps also play important roles by identifying relevant entities and structuring the partial response.

No Algorithm 1 ($\mathcal{A}_{MS}$) – Removing the mask search strategy results in slight declines in F1 and B4 (e.g. F1 drops by 0.3 on WoW Test Seen) but unexpected improvements in KF1 on OpenDialKG (+1.1 on Test Seen). This suggests that while $\mathcal{A}_{MS}$ refines mask placement for coherence, its absence may lead the model to focus on entity recall, boosting KF1 but sacrificing overall response quality.

F. CASE STUDY

We provide qualitative examples in Table 6 to evaluate the performance of DialFill and baseline methods on the WOW and OpenDialKG datasets. These examples highlight DialFill's ability to incorporate retrieved knowledge effectively and generate coherent responses.

In **Example 1** (WOW dataset), the baselines demonstrate varying levels of failure. DukeNet and KnowledGPT generate responses that are irrelevant to the user's query, with KnowledGPT suggesting "playing chess," which is unrelated to archery. GATE's response incorporates the correct topic (archery) but contains an illogical statement ("using a bow to move a bow"). In contrast, DialFill produces a relevant and accurate response that incorporates the retrieved knowledge ("using a bow to propel arrows"), directly addressing the user's question and maintaining fluency. This example highlights DialFill's ability to integrate knowledge while avoiding errors in expression or factuality.

In **Example 2** (WOW dataset), baseline methods exhibit significant factual inaccuracies. GATE misinterprets the timeline and incorrectly states that the individual was an undergraduate "in the 1800s." This type of hallucination undermines the credibility of the response. DialFill, on the

TABLE 6. Examples of model outputs on the WOW and OpenDialKG dataset, baselines' responses of Example 1 are reported from Sun et al. [13]. These examples show methods failing due to a text error (highlighted with red). DialFill generates a desired response that answers the user's question and incorporates retrived knowledge (highlighted with green).

Example 1: On the WOW dataset	
Context	User2: Huh. When did you start archery? Do you take part in any competitions? User1: I started when I was ten. I got a winning streak when I was 15, which means I won a consecutive sequence of games. User2: I think I want to enter one but I have no idea where to start.
Knowledge	Ground-truth: 2010–11 ISU Junior Grand Prix: The 2010–11 ISU Junior Grand Prix was the 14th season of the ISU Junior Grand Prix, a series of international junior level competitions organized by the International Skating Union. Retrived: Archery: Archery is the sport, practice or skill of using a bow to propel arrows.
Human	You could enter the Junior Grand Prix! It's a series of international junior level competitions.
Model outputs	
DukeNet	the earliest evidence of use traces to the olmecs (mexico) , with evidence of chocolate beverages dating to 1900 bc
KnowledGPT	you should start by playing chess. it's played by millions of people worldwide.
KnowExpert	i'm not sure, but i do know that archery is a sport that involves using a bow and arrow to shoot a long-range shot at targets on a moving target.
MSDP	I think I want to enter one but I have no idea where to start .
MixCL	You should start with the World Archery Championship. It's the biggest one in the world.
GATE	I think I would like to learn archery skills, like using a bow to move a bow.
DialFill	You can start by practicing using a bow to propel the arrows. I'm not sure where to start.
Example 2: On the WOW dataset	
Context	User1: How did he get involved in the music scene? model Wizard: He actually attended the Academy of Contemporary Music as an undergrad. User2: When did he first start gaining notoriety?
Knowledge	Ground-truth: Ed Sheeran: In early 2011, Sheeran independently released the extended play, "No." Retrived: Ed Sheeran: He attended the Academy of Contemporary Music in Guildford as an undergraduate from the age of 18 in 2009.
Human	He came out with No in 2011 and continued to gain fame since then.
Model outputs	
GATE	He was born in the 1800s. He was an undergraduate in the 1800s.
DialFill	He attended the Academy of Contemporary Music in Guildford as an undergraduate from 2009.
Example 3: On the OpenDialKG dataset	
Context	User1: Do you know Franz Kafka ? User2: He wrote Sämtliche Erzählungen and In der Strafkolonie . Have you heard of those before ?
Knowledge	Ground-truth: Sämtliche Erzählungen is written by John Updike Retrived: Sämtliche Erzählungen is written by John Updike"
Human	No , not yet ! What genre is Sämtliche Erzählungen ?
Model outputs	
DiffKG	I have not heard of them. I have heard of them. Is that the genre
NPH	I have not. Is he a writer?
GATE	I have not. Can you tell me more about them?
DialFill	No, I have not. Isn't Sämtliche Erzählungen written by John Updike?

other hand, accurately incorporates the retrieved knowledge (''He attended the Academy of Contemporary Music in Guildford as an undergraduate from 2009'') and delivers a coherent response. This showcases DialFill's capability to ground its responses in relevant and accurate knowledge, ensuring both contextual alignment and factual correctness.

In **Example 3** (OpenDialKG dataset), baseline models fail to utilize the retrieved knowledge effectively. For instance, GATE generates a generic response (''Can you tell me more about them?'') that neither incorporates the retrieved knowledge nor answers the user's query. In contrast, DialFill correctly incorporates the retrieved knowledge (''Sämtliche Erzählungen written by John Updike'') into its response, producing a relevant and knowledge-rich answer. This example demonstrates DialFill's ability to seamlessly integrate retrieved knowledge into its dialogue responses, aligning with user intent and maintaining coherence.

VIII. CONCLUSION

In this paper, we proposed DialFill, a Dialogue Filling (DF-based) framework designed to enhance the integration of external knowledge into dialogue systems. By reframing

response generation as a three-step process—keyword prediction, masked response generation, and dialogue filling—DialFill effectively bridges the gap between contextually relevant retrieval and coherent response generation. Extensive experiments on the Wizard-of-Wikipedia and OpenDialKG datasets demonstrate that DialFill significantly outperforms existing LM-based and KB-based methods in terms of relevance, factuality, and knowledge integration. Human evaluations and ablation studies further validate DialFill's ability to generate knowledge-base and contextually accurate responses.

While DialFill advances the state of knowledge-base dialogue systems, it has limitations, such as reliance on the quality of retrieved knowledge and occasional challenges in keyword prediction when knowledge is ambiguous. For future work, we aim to address these challenges by exploring more robust knowledge retrieval and selection strategies.

APPENDIX A
MASK SEARCH STRATEGY

This appendix provides a detailed explanation of the Mask Search Strategy employed in the *Masked Response Generation* step during inference (see Section V-B).

A. ALGORITHM

Algorithm 1 Mask Search Strategy

Input: Context x, Keyword $kw = \{kw_1, kw_2, \ldots, kw_m\}$, Maximum mask length $L_{\max}$ of each side
Output: Optimal mask lengths: $L_{\text{left}}, L_{\text{right}}$

Step 1: Optimize Left Mask Length
Initialize mask sequence $m = \{m_1, m_2, \ldots, m_{L_{\max}}\}$, where $m_i = [\text{mask}]$
Set target token $y_{\text{label}} = kw_1$
for $i = 1$ to $L_{\max}$ **do**
  Compute *score* for y_{label} given a left mask of length i:
    $score \leftarrow p_\theta(y_{\text{label}} \mid x, m_1, \ldots, m_i)$
end for
$L_{\text{left}} \leftarrow \arg\max_i score$

Step 2: Optimize Right Mask Length
Reinitialize mask sequence $m = \{m_1, m_2, \ldots, m_{L_{\max}}\}$
Set target token $y_{\text{label}} = [\text{EOS}]$
for $j = 1$ to $L_{\max}$ **do**
  Construct input sequence s=$(x, m_1, \ldots, m_{L_{\text{left}}}, kw, m_1, \ldots, m_j)$
  Compute *score* for y_{label} given s:
    $score \leftarrow p_\theta(y_{\text{label}} \mid s)$
end for
$L_{\text{right}} \leftarrow \arg\max_j score$
return $L_{\text{left}}, L_{\text{right}}$

B. ALGORITHM OVERVIEW

The Mask Search Strategy determines the optimal number of mask tokens to place on the left and right sides of the keyword kw in the masked response mr. This approach dynamically adjusts mask lengths to enhance the model's ability to generate coherent and contextually appropriate responses during the *Dialogue Filling* step.

1) STEP 1: OPTIMIZING LEFT MASK LENGTH

The algorithm initializes a sequence of mask tokens $m = \{m_1, m_2, \ldots, m_{L_{\max}}\}$, where each m_i represents a mask token. The target token y_{label} is set to the first token of the keyword kw_1.

For each potential mask length i from 1 to $L_{\max}$, the algorithm computes the probability *score* of generating y_{label} given the dialogue context x and a left mask of length i. The left mask length L_{left} that maximizes this probability is selected as the optimal length.

2) STEP 2: OPTIMIZING RIGHT MASK LENGTH

The algorithm then reinitializes the mask sequence for the right side and sets the target token y_{label} to the end-of-sequence token [EOS]. For each potential right mask length j from 1 to $L_{\max}$, it constructs the input sequence s by concatenating the context x, the optimized left mask, the keyword kw, and a right mask of length j.

The probability *score* of generating y_{label} given the sequence s is computed. The right mask length L_{right} that maximizes this probability is selected as the optimal length.

3) OUTCOME

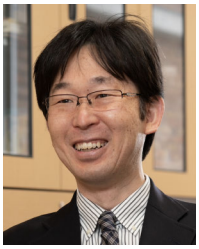
By optimizing L_{left} and L_{right}, the Mask Search Strategy ensures that the masked response mr facilitates effective *Dialogue Filling*. This method dynamically adjusts mask lengths based on the model's predictions, leading to improved response coherence and contextual relevance.

REFERENCES

[1] D. D. Freitas, M.-T. Luong, and D. R. So, "Towards a human-like open-domain Chatbot," 2020, *arXiv:2001.09977*.
[2] M. Huang, X. Zhu, and J. Gao, "Challenges in building intelligent open-domain dialog systems," *ACM Trans. Inf. Syst.*, vol. 38, no. 3, pp. 1–32, Jul. 2020.
[3] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.
[4] A. Grattafiori et al., "The llama 3 herd of models," 2024, *arXiv:2407.21783*.
[5] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval augmentation reduces hallucination in conversation," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*, 2021, pp. 3784–3803.
[6] Z. Ji, N. Lee, and R. Frieske, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, 2022.
[7] S. Moon, P. Shah, A. Kumar, and R. Subba, "OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 845–854. [Online]. Available: https://aclanthology.org/P19-1081
[8] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, "Wizard of wikipedia: Knowledge-powered conversational agents," 2018, *arXiv:1811.01241*.
[9] M. Ghazvininejad, C. Brockett, M. Chang, B. Dolan, J. Gao, W.-T. Yih, and M. Galley, "A knowledge-grounded neural conversation model," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, Apr. 2018, pp. 5110–5117. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11977
[10] Z. Liu, M. Patwary, R. Prenger, S. Prabhumoye, W. Ping, M. Shoeybi, and B. Catanzaro, "Multi-stage prompting for knowledgeable dialogue generation," in *Proc. Findings Assoc. Comput. Linguistics (ACL)*, 2022, pp. 1317–1337.
[11] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E. M. Smith, Y.-L. Boureau, and J. Weston, "Recipes for building an open-domain chatbot," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics, Main Volume*, 2021, pp. 300–325.

[12] R. Thoppilan et al., "LaMDA: Language models for dialog applications," 2022, *arXiv:2201.08239*.
[13] W. Sun, Z. Shi, S. Gao, P. Ren, M. de Rijke, and Z. Ren, "Contrastive learning reduces hallucination in conversations," 2022, *arXiv:2212.10400*.
[14] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-T. Yih, "Dense passage retrieval for open-domain question answering," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 5835–5847.
[15] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," 2020, *arXiv:2005.11401*.
[16] S. Cao and L. Wang, "CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization," 2021, *arXiv:2109.09209*.
[17] L. Qin, Y. Zhang, H. Liang, J. Wang, and Z. Yang, "Well begun is half done: Generator-agnostic knowledge pre-selection for knowledge-grounded dialogue," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Singapore: Association for Computational Linguistics, 2023, pp. 4696–4709. [Online]. Available: https://aclanthology.org/2023.emnlp-main.285
[18] Y. Xu, E. Ishii, S. Cahyawijaya, Z. Liu, G. I. Winata, A. Madotto, D. Su, and P. Fung, "Retrieval-free knowledge-grounded dialogue response generation with adapters," in *Proc. 2nd DialDoc Workshop Document-Grounded Dialogue Conversational Question Answering*, 2022, pp. 93–107.
[19] Z. Wu, W. Bi, X. Li, L. Kong, and B. Kao, "Lexical knowledge internalization for neural dialog generation," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2022, pp. 7945–7958.
[20] X. Zhao, W. Wu, C. Xu, C. Tao, D. Zhao, and R. Yan, "Knowledge-grounded dialogue generation with pre-trained language models," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 3377–3390.
[21] C. Zheng and M. Huang, "Exploring prompt-based few-shot learning for grounded dialog generation," 2021, *arXiv:2109.06513*.
[22] C. Donahue, M. Lee, and P. Liang, "Enabling language models to fill in the blanks," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2492–2501. [Online]. Available: https://aclanthology.org/2020.acl-main.225
[23] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., (Long Short Papers)*, vol. 1, Jan. 2018, pp. 4171–4186.
[24] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 5485–5551, Jan. 2019.
[25] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang, "GLM: General language model pretraining with autoregressive blank infilling," 2021, *arXiv:2103.10360*.
[26] Y. Zhao, W. Wu, and C. Xu, "Are pre-trained language models knowledge-able to ground open domain dialogues?" 2020, *arXiv:2011.09708*.
[27] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. ACL*, Jul. 2004, pp. 74–81.
[28] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*. Philadelphia, PA, USA: Association for Computational Linguistics, 2002, pp. 311–318.
[29] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proc. 9th Workshop Stat. Mach. Transl.*, 2014, pp. 376–380.
[30] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-LM: Training multi-billion parameter language models using model parallelism," 2019, *arXiv:1909.08053*.
[31] Y.-L. Tuan, S. Beygi, M. Fazel-Zarandi, Q. Gao, A. Cervone, and W. Y. Wang, "Towards large-scale interpretable knowledge graph reasoning for dialogue systems," in *Proc. Findings Assoc. Comput. Linguistics (ACL)*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 383–395. [Online]. Available: https://aclanthology.org/2022.findings-acl.33
[32] N. Dziri, A. Madotto, O. Zaïane, and A. J. Bose, "Neural path hunter: Reducing hallucination in dialogue systems via path grounding," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 2197–2214. [Online]. Available: https://aclanthology.org/2021.emnlp-main.168
[33] C. Meng, P. Ren, Z. Chen, W. Sun, Z. Ren, Z. Tu, and M. D. Rijke, "DukeNet: A dual knowledge interaction network for knowledge-grounded conversation," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 1151–1160.
[34] J. Bai, Z. Yang, J. Yang, H. Guo, and Z. Li, "KINet: Incorporating relevant facts into knowledge-grounded dialog generation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 1213–1222, 2023.
[35] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," 2021, *arXiv:2106.09685*.
[36] Q. Xue, T. Takiguchi, and Y. Ariki, "Building a knowledge-based dialogue system with text infilling," in *Proc. 23rd Annu. Meeting Special Interest Group Discourse Dialogue*, 2022, pp. 237–243. [Online]. Available: https://aclanthology.org/2022.sigdial-1.25

QIANG XUE received the B.Eng. degree from Dalian Jiaotong University, China, in 2016, and the M.Eng. degree from Kobe University, Japan, in 2021. His research interests include natural language processing and machine learning applied to dialogue systems.



TETSUYA TAKIGUCHI (Member, IEEE) received the M.Eng. and Dr.-Eng. degrees in information science. He was a Researcher with Tokyo Research Laboratory, IBM Research. From 2004 to 2016, he was an Associate Professor with Kobe University, where he has been Professor, since 2016. From May 2008 to September 2008, he was a Visiting Scholar with the Department of Electrical Engineering, University of Washington. From March 2010 to September 2010, he was a Visiting Scholar with the Institute for Learning and Brain Sciences, University of Washington. From April 2013 to October 2013, he was a Visiting Scholar with the Laboratoire d'InfoRmatique en Image et Systèmes d'information, INSALyon. His research interests include speech, image, and brain processing, and multimodal assistive technologies for people with articulation disorders. He is currently a member of IEICE, IPSJ, and ASJ.

YASUO ARIKI (Life Member, IEEE) received the B.E., M.E., and Ph.D. degrees in information science from Kyoto University, in 1974, 1976, and 1979, respectively. He was an Assistant Professor with Kyoto University, from 1980 to 1990, and a Visiting Academician with The University of Edinburgh, from 1987 to 1990. He was an Associate Professor and a Professor with Ryukoku University, Japan, from 1990 to 1992 and from 1992 to 2003, respectively. From 2003 to 2016, he was a Professor with Kobe University. He is mainly engaged in speech and image recognition and is interested in dialogue systems. He is a member of IEICE, IPSJ, JSAI, ASJ, and NLP.