

# Efficient Tuning of Large Language Models for Knowledge-Grounded Dialogue Generation

Bo Zhang<sup>1</sup> Hui Ma<sup>2</sup> Dailin Li<sup>1</sup> Jian Ding<sup>1</sup> Jian Wang<sup>1\*</sup>  
Bo Xu<sup>1</sup> HongFei Lin<sup>1</sup>

<sup>1</sup>Dalian University of Technology, China <sup>2</sup>Hefei University of Technology, China  
{zhangbo1998, ldlbest, 91mr\_ding}@mail.dlut.edu.cn  
huima@hfut.edu.cn, {wangjian, xubo, hflin}@dlut.edu.cn

## Abstract

Large language models (LLMs) demonstrate remarkable text comprehension and generation capabilities but often lack the ability to utilize up-to-date or domain-specific knowledge not included in their training data. To address this gap, we introduce KEDiT, an efficient method for fine-tuning LLMs for knowledge-grounded dialogue generation. KEDiT operates in two main phases. First, it employs an information bottleneck to compress retrieved knowledge into learnable parameters, retaining essential information while minimizing computational overhead. Second, a lightweight knowledge-aware adapter integrates these compressed knowledge vectors into the LLM during fine-tuning, updating less than 2% of the model parameters. The experimental results on the Wizard of Wikipedia and a newly constructed PubMed-Dialog dataset demonstrate that KEDiT excels in generating contextually relevant and informative responses, outperforming competitive baselines in automatic, LLM-based, and human evaluations. This approach effectively combines the strengths of pretrained LLMs with the adaptability needed for incorporating dynamic knowledge, presenting a scalable solution for fields such as medicine.<sup>1</sup>

## 1 Introduction

The field of natural language processing has undergone a significant transformation recently with the advent of large language models (LLMs) (Brown et al., 2020; OpenAI, 2022, 2023; Touvron et al., 2023; Chowdhery et al., 2023). These models, characterized by their vast number of parameters, have demonstrated remarkable abilities in understanding and generating human-like text,

powered by extensive pretraining on diverse and extensive datasets. However, LLMs struggle with tasks that require up-to-date knowledge or domain-specific expertise that was not included in their training datasets (Kandpal et al., 2023; Zhang et al., 2023b). This limitation has led to the exploration of methods to augment LLMs with external knowledge, thereby improving their performance in knowledge-intensive tasks.

One promising approach to address this challenge is retrieval-augmented generation (RAG), a technique that integrates retrieval mechanisms into the generative process of LLMs (Lewis et al., 2020b; Guu et al., 2020; Borgeaud et al., 2022; Siriwardhana et al., 2023; Ram et al., 2023; Izacard et al., 2023). This method allows LLMs to access and utilize external, relevant information dynamically, as they generate responses. Current methods in the RAG system typically employ off-the-shelf LLMs combined with general-purpose retrievers, leveraging the inherent in-context learning capabilities of these language models (Ram et al., 2023; Yu et al., 2023; Sarthi et al., 2024). However, this approach encounters limitations when the LLMs are not specifically trained to incorporate retrieved content, particularly in utilizing domain-specific knowledge. These challenges are exacerbated in fields where accurate, specialized information is crucial. In contrast, other researchers have adopted an end-to-end training approach, integrating both LLMs and retrieval mechanisms (Lin et al., 2024; Asai et al., 2024). This method undoubtedly improves the overall performance of the system by aligning the learning objectives of the model with the retrieval tasks directly. Nonetheless, these extensive training processes are resource-intensive and costly, posing significant challenges for deploying these models in environments that demand up-to-date knowledge integration. Given these challenges, we shift our focus to the generative

\* The corresponding author.

<sup>1</sup>Code and data are available at: <https://github.com/zhangbo-nlp/KEDiT>.

aspect to enhance knowledge-grounded dialogue generation directly.

Unlike previous approaches to knowledge-grounded dialogue generation (Meng et al., 2021; Zhang et al., 2022; Xu et al., 2023), which often involve a knowledge selection step, we do not consider this step because of its computational expense when it is applied to LLMs, such as with LSR (Shi et al., 2024). Instead, we propose an efficient method for fine-tuning LLMs for knowledge-grounded dialogue generation, named KEDiT. This method directly utilizes the retrieved knowledge without the selection process and consists of two main stages. **First**, we employ an information bottleneck method (Li et al., 2023; Zhang et al., 2025) to compress the retrieved knowledge into a set of learnable vectors by maximizing the mutual information between the original knowledge and these compressed vectors. This approach ensures that essential information is retained and reduces the computational complexity of processing extensive knowledge inputs. To further improve this representation, we introduce an alignment loss, refining the compressed vectors to align them with the internal representations of the LLM. **Second**, we integrate the compressed knowledge representation into the dialogue generation process through a lightweight knowledge-aware adapter (KA-Adapter), which improves the model by inserting small, trainable modules into its architecture. These modules, integrated into both the attention layer and feed-forward layer, selectively fine-tune the model while keeping the majority of its parameters frozen. The adapter employs a gating mechanism as a control channel, regulating how the compressed vectors influence the internal states of the LLM. This design ensures that external knowledge is effectively incorporated without disrupting the pretrained representations. Requiring fine-tuning of less than 2% of the model parameters, the KA-Adapter balances computational efficiency and high performance in knowledge-grounded dialogue tasks.

To validate the effectiveness of KEDiT, we conduct comprehensive experiments in both open-domain and specialized domain settings. For the open-domain evaluation, we utilize the Wizard of Wikipedia dataset (Dinan et al., 2019) to test the ability of models to generate responses grounded in a wide range of knowledge topics. Additionally, to assess performance in specialized do-

main and with up-to-date information, we create a domain-specific dialogue dataset using GPT-4o, based on the latest research from PubMed.<sup>2</sup> The experimental results demonstrate that KEDiT achieves substantial improvements over competitive baselines in automatic, LLM-based, and human evaluations. KEDiT shows superior performance in generating contextually relevant and informative responses and excels in handling domain-specific knowledge. To further validate the practicality of the proposed method, we evaluate KEDiT across various domains and tasks, highlighting its adaptability and robustness in diverse scenarios.

In summary, we present KEDiT, an efficient method for improving knowledge-grounded dialogue generation in large language models. By integrating knowledge-aware components, KEDiT offers a scalable solution to the challenge of incorporating extensive and evolving knowledge bases into dialogue systems without extensive costs. Furthermore, we introduce a new domain-specific dialogue dataset, PubMed-Dialog, which provides a benchmark for assessing the ability of the model to address specialized, up-to-date biomedical information in dialogue generation.

## 2 Related Work

### 2.1 Knowledge-Grounded Dialogue

Knowledge-grounded dialogue systems generate responses that are both contextually appropriate and enriched with relevant information drawn from external knowledge sources. Traditional methods involve pretrained language models such as BART (Lewis et al., 2020a) and T5 (Raffel et al., 2020), which are fine-tuned on dialogue datasets with explicit knowledge selection and integration (Dinan et al., 2019; Kim et al., 2020; Meng et al., 2021; Shen et al., 2022). These methods typically use a two-step process: selecting relevant knowledge and generating responses on the basis of this information. For example, certain models, such as SPI (Xu et al., 2023), select the single most relevant knowledge sentence for generation, whereas other models, such as TransIKG (Zhang et al., 2022), integrate multiple knowledge sentences using mechanisms like attention. However, these methods assume that the gold standard knowledge is contained within the available knowledge base,

<sup>2</sup><https://pubmed.ncbi.nlm.nih.gov/>.

which limits their generalizability. In contrast, Liu et al. (2021) proposed a weakly supervised learning framework, and Bai et al. (2023) introduced knowledgeable prefix tuning to inject all relevant knowledge directly into the model, bypassing the need for knowledge selection. However, these innovative methods cannot be applied to LLMs directly because of their specific architectures.

The advent of RAG introduced a new dimension to knowledge-grounded dialogue by combining retrieval mechanisms with LLMs, enabling them to dynamically access external information (Lewis et al., 2020b; Guu et al., 2020; Borgeaud et al., 2022; Siriwardhana et al., 2023; Ram et al., 2023; Izacard et al., 2023). For example, Ram et al. (2023) show that retrieval-augmented language modeling significantly improves performance by conditioning on relevant documents without modifying the language model. These methods increase the flexibility and applicability of LLMs in knowledge-intensive tasks, but they still struggle to utilize new knowledge in specialized domains efficiently, because they are not specifically trained to incorporate retrieved content. Recent approaches, such as Lin et al. (2024), have explored end-to-end training of LLMs with integrated retrieval mechanisms to address these limitations. However, these methods are resource-intensive and challenging to deploy in environments requiring frequent updates.

Our method, KEDiT, addresses these challenges by compressing retrieved knowledge into learnable parameters and integrating them through a lightweight adapter. This approach reduces computational overhead and maintains high performance and adaptability in both open-domain and specialized domain settings.

## 2.2 Parameter-Efficient Fine-Tuning

Parameter-efficient fine-tuning techniques adapt LLMs to new tasks with minimal parameter updates, reducing computational costs. These methods can be broadly categorized into two types: adapter-based and prompt-based methods.

Adapter-based methods, such as those introduced by Houshy et al. (2019) and extended by others (Karimi Mahabadi et al., 2021; Hu et al., 2022; Zhang et al., 2023a, 2024), insert additional trainable parameters into the model architecture while keeping most of the original model weights frozen. Among these, low-rank adaptation (LoRA)

(Hu et al., 2022) has emerged as a prominent technique that employs a pair of smaller matrices to update the model weights through low-rank decomposition. Prompt-based methods offer another approach to parameter-efficient fine-tuning. These methods, including prompt tuning (Lester et al., 2021), prefix tuning (Li and Liang, 2021), and P-tuning (Liu et al., 2024, 2022), prime a frozen pretrained model for a downstream task by including a trainable collection of tokens either in the input embeddings or at every intermediate layer of the model. Additionally, He et al. (2022) combined prefix tuning and LoRA to propose the MAM Adapter, which further refines parameter-efficient fine-tuning by applying modifications to specific hidden states in the model. However, these methods face challenges in knowledge-intensive tasks where specialized information is crucial, often failing to fully capture and utilize the complexity of this knowledge.

Recent parameter-efficient approaches for knowledge-grounded tasks include KnowExpert (Xu et al., 2022), which utilizes specialized adapters encoding fixed topic-specific information, and KnowPrefix-Tuning (Bai et al., 2023), which employs knowledge prefixes to prompt latent information. However, both methods are limited by their reliance on knowledge encoded during pre-training or fixed during training.

Our proposed KEDiT is specifically designed for knowledge-grounded tasks, focusing on reducing computational overhead while effectively leveraging external dynamic knowledge. Furthermore, we innovatively adapt the prompt-based method into an adapter-based approach with a gating mechanism, ensuring seamless integration of compressed knowledge vectors into the LLM for improved dialogue generation.

## 3 Method

### 3.1 Task Statement and Model Overview

Given a dialogue context  $C$  and a set of retrieved knowledge pieces  $K = \{k_1, k_2, \dots, k_n\}$ , the task is to generate a response  $R$  that is both contextually appropriate and enriched with the provided knowledge. Formally, we aim to maximize the conditional probability  $p(R|C, K)$ . However, using  $K$  directly can be computationally expensive for LLMs. To address this issue, we propose KEDiT, as shown in Figure 1a, which comprises two main components: a knowledge bottleneck

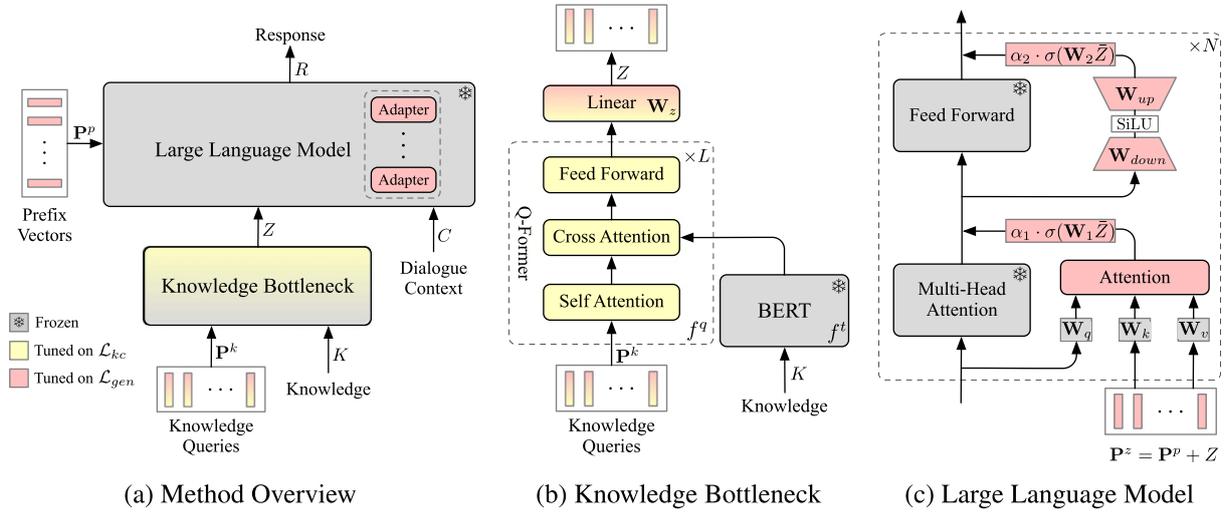


Figure 1: (a) Overall architecture of KEDiT, showing the flow from the input knowledge and dialogue context to the generated response. (b) Detailed structure of the knowledge bottleneck module, showing how BERT and the Q-Former compress knowledge into a compact representation through multi-head attention and feed-forward layers. (c) Integration of compressed knowledge into the large language model using KA-Adapter, detailing KA-Attn and KA-FFN. For simplicity, we omit adding & norm layers to the diagrams. Yellow indicates tuning in the knowledge compression step, pink indicates tuning in the dialogue generation step, and gray represents frozen modules.

module and a frozen LLM enhanced with KA-Adapter. The knowledge bottleneck module distills essential information from  $K$  into a compact representation  $Z$ , formalized by  $p_\phi(Z|K)$ , where  $Z$  is a set of learnable vectors. The LLM then incorporates  $Z$  into the dialogue generation process via the KA-Adapter, represented by  $p_\theta(R|C, Z)$ . Thus, our revised objective becomes:

$$p(R|C, K) \approx p_\theta(R|C, Z)p_\phi(Z|K). \quad (1)$$

In summary, our approach follows a two-step strategy: knowledge compression to distill  $K$  into  $Z$  and knowledge integration to efficiently utilize  $Z$  within the dialogue generation process. The training and inference procedures are described in the following sections, with an overview provided in Appendix A.

### 3.2 Knowledge Compression via an Information Bottleneck

Integrating external knowledge into LLMs presents significant challenges due to the vast length of retrieved information, which often contains irrelevant details and increases computational costs. To address this issue, we propose a knowledge compression mechanism based on the information bottleneck principle. This method distills essential information into a fixed-size, learnable

representation, balancing relevance and efficiency for seamless integration with LLMs. Our approach trains a knowledge bottleneck module, combining BERT (Devlin et al., 2019) and the Q-Former (Li et al., 2023), that is optimized using an information bottleneck objective on a large-scale knowledge dataset  $\mathcal{D}_k$ .

**Knowledge Encoding** We utilize a pretrained BERT to encode the knowledge  $K$  into feature representations  $f^t(K)$ , specifically using the last hidden states. These features are then processed by the Q-Former, which compresses them into a compact representation  $Z$ . As shown in Figure 1b, the Q-Former consists of  $L$  blocks, each including a self-attention layer, followed by a cross-attention layer that incorporates  $f^t(K)$ , and a feed-forward network. Each block is formally defined as:

$$\mathbf{Z}_\ell = f_\ell^q(\mathbf{Z}_{\ell-1}, f^t(K)) \in \mathbb{R}^{m \times d_1}, \quad (2)$$

for  $\ell = 1, \dots, L$ , with the initial input  $\mathbf{Z}_0 = \mathbf{P}^k$ , where  $\mathbf{P}^k \in \mathbb{R}^{m \times d_1}$  represents  $m$  learnable vectors referred to as knowledge queries. These vectors are randomly initialized and interact with  $f^t(K)$ , which enables them to absorb the semantic and contextual information from  $K$ . The concept of knowledge queries is inspired by prompt tuning

(Lester et al., 2021). They serve as trainable vectors designed to capture essential information from  $K$  and integrate it efficiently into the LLM.

After processing through all  $L$  blocks, we obtain the final representation  $Z = \mathbf{W}_z \mathbf{Z}_L \in \mathbb{R}^{m \times d_z}$ , where  $\mathbf{W}_z$  is a learnable projection matrix that maps the representation to the dimension required by the LLM.

**Mutual Information Optimization** To ensure that the compressed  $Z$  retains the most essential information from  $K$  and can be fully utilized by the LLM, we maximize their mutual information  $I(K; Z)$  via the LLM. A common approach for achieving this is to maximize a variational lower bound (Barber and Agakov, 2003) on the mutual information, which is expressed as:

$$I(K; Z) \geq \mathbb{E}_{p(K, Z)} \log q(K|Z) + H(K). \quad (3)$$

However, in our setup,  $Z$  is a set of learnable vectors rather than latent variables sampled from a specific distribution. Therefore, we adapt this approach by introducing an auxiliary model  $q_\psi(K|Z)$  to reconstruct  $K$  from  $Z$ . This model is parameterized by the LLM, thereby ensuring that  $Z$  can be effectively utilized by the LLM in the dialogue generation process. Consequently, our optimization objective simplifies to:

$$\mathcal{L}_{recon} = -\mathbb{E}_{p(K)p_\phi(Z|K)} \log q_\psi(K|Z), \quad (4)$$

where  $p_\phi(Z|K)$  is modeled by the knowledge bottleneck. This formulation is similar to the variational lower bound approach, where  $p(K, Z)$  is factorized as  $p(K) \times p_\phi(Z|K)$  and the entropy term  $H(K)$ , which is constant with respect to the model parameters, can be omitted during optimization. Although this adaptation diverges from traditional variational approaches, we find that it works well in practice.

**Alignment Loss** While mutual information optimization ensures that  $Z$  retains essential information from  $K$ , it does not guarantee that  $Z$  is readily interpretable by the LLM. To further refine the compressed knowledge vectors  $Z$  and align them closely with the LLM’s internal representations, we introduce an alignment loss. This loss is designed to ensure that the compressed knowledge vectors are easily interpretable and utilizable by the LLM. Specifically, alignment

loss ensures that the structure of the compressed vectors  $Z$  produced by the knowledge bottleneck is compatible with the vectors  $\hat{Z}$  reconstructed by the LLM from the original knowledge, where  $\hat{Z}$  corresponds to the final hidden states of the LLM. We define this loss as the mean squared error between the vectors  $Z \sim p_\phi(Z|K)$  and the vectors  $\hat{Z} \sim q_\psi(Z|K)$ :

$$\mathcal{L}_{align} = \frac{1}{m} \sum_{i=1}^m (Z_i - \hat{Z}_i)^2. \quad (5)$$

**Training Objective** The overall objective for knowledge compression combines mutual information optimization and alignment loss:

$$\mathcal{L}_{kc} = \mathcal{L}_{recon} + \beta \cdot \mathcal{L}_{align}, \quad (6)$$

where  $\beta$  is a hyperparameter that balances the contribution of the alignment loss. During this training phase, the parameters of the LLM and BERT are kept frozen, whereas only the parameters of the Q-Former are trained.

By minimizing this objective, we ensure that the compressed knowledge vectors  $Z$  capture essential information from  $K$  effectively while being well-aligned with the internal representations of the LLM. This compressed representation is then used in the dialogue generation process, as described in the subsequent sections.

### 3.3 Knowledge Integration into Dialogue Generation

After the compressed knowledge representation  $Z$  is obtained, the next step is to integrate this knowledge into the LLM to improve dialogue generation while maintaining computational efficiency. Directly fine-tuning the LLM to incorporate external knowledge is both resource-intensive and risks disrupting the pretrained representations of the model. To address this issue, we design the KA-Adapter, a lightweight module that integrates external knowledge by inserting small, trainable components into the LLM layers. This approach preserves the pretrained capabilities of the model and enables efficient fine-tuning on a knowledge-grounded dialogue dataset  $\mathcal{D}_d$ . As depicted in Figure 1c, this adapter consists of two main components: the knowledge-aware attention mechanism (KA-Attn) and the knowledge-aware feed-forward network (KA-FFN).

**Knowledge-Aware Attention Mechanism** KA-Attn improves the standard self-attention mechanism of LLMs by incorporating  $Z$ . Inspired by an alternative view of prefix tuning (He et al., 2022), we transform prompt-based prefix tuning into an adapter-based method, adapting the attention mechanism as follows:

$$\begin{aligned} \mathbf{h}_1 &\leftarrow \mathbf{h}_1 + \alpha_1 \cdot \sigma(\mathbf{W}_1 \bar{Z}) \cdot \Delta \mathbf{h}_1, \\ \Delta \mathbf{h}_1 &= \text{Softmax}(\mathbf{W}_q \mathbf{x} (\mathbf{W}_k \mathbf{P}^z)^T) \mathbf{W}_v \mathbf{P}^z, \end{aligned} \quad (7)$$

where  $\mathbf{P}^z = \mathbf{P}^p + Z$ ,  $\mathbf{P}^p \in \mathbb{R}^{m \times d_2}$  represents  $m$  learnable prefix vectors,  $\mathbf{x}$  is the input feature, and  $\sigma$  denotes the *sigmoid* activation.  $\mathbf{W}_q$ ,  $\mathbf{W}_k$ , and  $\mathbf{W}_v$  are the query, key, and value matrices, respectively, of the original attention mechanism. The function  $\alpha_1 \cdot \sigma(\mathbf{W}_1 \bar{Z})$  acts as a gating mechanism, allowing the model to dynamically adjust its focus on the adaptive change on the basis of the mean value of the compressed vectors  $Z$ .

**Knowledge-Aware Feed-Forward Network** Similarly, the KA-FFN is designed to integrate the compressed knowledge vectors into the standard FFN layer of LLMs. This mechanism draws inspiration from LoRA and a scaled parallel adapter (He et al., 2022) and is adapted to better integrate external knowledge as follows:

$$\begin{aligned} \mathbf{h}_2 &\leftarrow \mathbf{h}_2 + \alpha_2 \cdot \sigma(\mathbf{W}_2 \bar{Z}) \cdot \Delta \mathbf{h}_2, \\ \Delta \mathbf{h}_2 &= \mathbf{W}_{up} \cdot \text{SiLU}(\mathbf{W}_{down} \mathbf{h}_1), \end{aligned} \quad (8)$$

where  $\mathbf{W}_{down}$  and  $\mathbf{W}_{up}$  are matrices that transform  $\mathbf{h}_1$  into a lower and then back to a higher dimension, effectively performing a bottleneck operation. The term  $\alpha_2 \cdot \sigma(\mathbf{W}_2 \bar{Z})$  functions similarly to the gating mechanism in KA-Attn, modulating the impact of  $\Delta \mathbf{h}_2$  on the FFN’s output.

**Generating Response** To generate the response  $R$ , the LLM predicts each token  $R_t$  sequentially, conditioned on the dialogue context  $C$ , the compressed representation  $Z$ , and the previously generated tokens  $R_{<t}$ . At each step, the input embeddings of  $C$  and  $R_{<t}$  are concatenated with  $Z$  and then passed into the LLM’s stacked  $N$  layers with the KA-Adapter. The final hidden state  $\mathbf{h}_{R_t}$  of the current token is then projected into the vocabulary space to compute the next-token probabilities:

$$p_\theta(R_t | R_{<t}, C, Z) = \text{Softmax}(\mathbf{W}_o \mathbf{h}_{R_t}), \quad (9)$$

where  $\mathbf{W}_o$  is the output projection matrix.

**Training Objective** The training objective for this stage focuses on ensuring that the LLM generates contextually appropriate and knowledge-enriched responses. We achieve this by minimizing the negative log-likelihood of the target response tokens:

$$\mathcal{L}_{gen} = - \sum_t \log p_\theta(R_t | R_{<t}, C, Z), \quad (10)$$

where  $(C, K, R) \in \mathcal{D}_d$  and  $Z \sim p_\phi(Z | K)$ . In this step, only the parameters of the KA-Adapter and the projection matrix  $\mathbf{W}_z$  are tuned, whereas the parameters of the LLM and the remainder of the knowledge bottleneck remain frozen.

## 4 Experiments

### 4.1 Datasets

To evaluate the effectiveness of KEDiT, we employ three datasets. The Wikipedia dataset serves as the knowledge corpus ( $\mathcal{D}_k$ ) for knowledge compression training. The Wizard of Wikipedia and PubMed-Dialog datasets are used for training and evaluating dialogue generation ( $\mathcal{D}_d$ ).

**Wikipedia** For the knowledge compression phase, we use the English version of the Wikipedia dataset.<sup>3</sup> This dataset is built from Wikipedia dumps and includes cleaned articles. We further process this dataset by splitting articles into paragraphs, selecting up to 500 words of content per article, and discarding articles with fewer than 50 words. This process results in a high-quality dataset of 6 million text chunks, which is suitable for training our knowledge bottleneck model.

**Wizard of Wikipedia** This dataset (Dinan et al., 2019) is used to evaluate the model performance in open-domain dialogue generation. The dataset contains 22.3k dialogues with 100.8k turns, where the agent provides informative responses grounded in Wikipedia knowledge. The dataset is divided into training, validation, and test sets, with the validation and test sets further split into seen and unseen categories. The seen category includes topics present in the training set, whereas the unseen category consists of dialogues with topics not encountered during training. In our experiments, instead of using the predefined knowledge sentences provided in the dataset, we

<sup>3</sup><https://huggingface.co/datasets/wikimedia/wikipedia>.

retrieve three relevant knowledge pieces from the provided knowledge topic articles on the basis of the dialogue context using TF-IDF, as described in Dinan et al. (2019). Notably, only 80% of the dialogue turns contain the predefined gold knowledge sentence among the retrieved topics. Although this may slightly impact performance, it better simulates real-world scenarios.

**PubMed-Dialog** To evaluate the model performance in specialized domains, we construct the PubMed-Dialog dataset using GPT-4o. Specifically, we use a data filtering methodology similar to that used in PubMedQA (Jin et al., 2019) to select relevant latest research articles from the PubMed database. Next, we design a prompt to instruct GPT-4o to generate multi-turn dialogues on the basis of the knowledge from the corresponding abstracts of these articles. This prompt is crafted to simulate natural conversations about biomedical topics, ensuring that the dialogues cover a range of aspects related to the topics discussed in the abstract. This provides a comprehensive benchmark for assessing the ability of the model to address specialized, up-to-date knowledge in dialogue generation. Notably, we directly use the abstract as knowledge of the corresponding dialogue context. To ensure the quality and faithfulness of the generated dialogues and minimize hallucinations, where content may deviate from the source information or contain inaccuracies, we implement a multi-iterative validation process. This process involves three rounds of evaluation and regeneration, iteratively refining the dialogues until they meet the required standards of consistency and factual correctness. Following this process, we obtain a dataset of 10.9k dialogues, with each dialogue containing an average of 4.36 turns. The dataset is divided into training, validation, and test sets, with 80% for training, 10% for validation, and 10% for testing. Details on the prompt design and the multi-iterative validation process are provided in Appendix B.

## 4.2 Experimental Setup

**Baseline Models** To evaluate the performance of KEDiT, we compare it against several baseline models grouped into three categories. First, the standard language models, including BART-Large (Lewis et al., 2020a) and Llama-3-8B (Touvron et al., 2023) using LoRA (LLAMA<sub>lora</sub>), generate dialogue responses without external knowledge,

which serve as basic benchmarks. Second, traditional knowledge-grounded models, such as TransIKG (Zhang et al., 2022) and SPI (Xu et al., 2023), improve dialogue generation by selecting and incorporating predefined knowledge sentences from fixed sources in the original dataset. Third, the retrieved knowledge-augmented models, such as KAT-TSLF (Liu et al., 2021), Llama-3-8B using KnowPrefix-Tuning (Bai et al., 2023) (LLAMA<sub>kpt</sub>), and Llama-3-8B using retrieved knowledge for zero-shot generation (LLAMA<sub>rag</sub>), utilize knowledge pieces retrieved on the basis of the dialogue history using the same TF-IDF method as in our experiments. We apply the same data processing methods within each category. All baselines are fine-tuned on the target dialogue dataset for fair comparison.

**Implementation Details** In our implementation of KEDiT, we utilize Llama-3-8B as the frozen LLM. Both the BERT encoder and the Q-Former are initialized with weights from BERT<sub>base</sub> (Devlin et al., 2019). During the knowledge compression phase, the knowledge bottleneck is trained on  $\mathcal{D}_k$  using the AdamW optimizer with a batch size of 64 over 1 epoch. The training configuration includes a learning rate of  $2e-4$ , 10,000 warmup steps, a weight decay of 0.05, and  $\beta$  is 0.5. The parameter  $m$  is 16 to balance the trade-off between the expressiveness of the compressed knowledge representation and computational efficiency. The impact of different values of  $m$  is discussed in Section 4.5.4. The models  $q_\psi(Z|K)$  and  $q_\psi(K|Z)$  are generated using instructions detailed in Appendix D. In the dialogue generation phase, the KA-Adapter is fine-tuned on  $\mathcal{D}_d$  using the AdamW optimizer with a learning rate of  $1e-4$  and cosine decay. Training is conducted over 3 epochs with a batch size of 32, and  $\alpha_1$  and  $\alpha_2$  are 2 and 4, respectively. The generation model  $p_\theta(R|C, Z)$  uses instruction templates specified in Appendix D to concatenate  $C$  and  $Z$ . We concatenate all knowledge snippets collectively and input them into BERT for encoding.

## 4.3 Evaluation Methods

**Automatic Evaluation** To quantify the performance of KEDiT, we utilize the following metrics: (1) BLEU, which measures the precision of  $n$ -grams in the generated responses compared with reference responses, with the final score being the average of BLEU-1, BLEU-2,

BLEU-3, and BLEU-4; (2) ROUGE, which evaluates the recall of  $n$ -grams, focusing on ROUGE-1, ROUGE-2, and ROUGE-L, with the final score being the average of these three metrics; and (3) F1 score, specifically the unigram F1 score, which is the harmonic mean of precision and recall for unigrams.

**LLM-Based Evaluation** Inspired by the evaluation framework in Zheng et al. (2023), we use GPT-4o as an advanced judge to assess the quality of the responses of our model. We employ two key methods: pairwise comparison, where GPT-4o is presented with a user query and two responses (one response from a baseline and one response from KEDiT) and selects the better response on the basis of relevance, informativeness, accuracy, and coherence; and multi-response grading, where GPT-4o is presented with responses from several baselines and KEDiT simultaneously and assigns scores from 1 to 5 for relevance, informativeness, and fluency for each response. This method has demonstrated high agreement with human evaluations, exceeding 80% alignment, comparable to human-human agreement levels (Zheng et al., 2023). For a detailed evaluation, we randomly sample 500 examples each from the test seen and test unseen sets of the Wizard of Wikipedia dataset, and 1000 examples from the PubMed-Dialog test set. Specific prompts for these evaluations are provided in Appendix C.

**Human Evaluation** To complement the LLM-based evaluation and validate its reliability, we conduct a human evaluation as an additional assessment of the model performance. Specifically, we randomly sample 20% of the data from the LLM-based evaluation sets: 100 examples each from the test seen and test unseen sets of the Wizard of Wikipedia dataset, and 200 examples from the PubMed-Dialog test set, totaling 400 samples. We recruited 12 graduate students with expertise in natural language processing and bioinformatics to serve as evaluators. These evaluators are divided into four groups of three, with each group independently evaluating 100 samples. Evaluators assess responses on the basis of relevance, informativeness, and fluency using a 1-to-5 Likert scale, following the same criteria as the LLM-based evaluation. To ensure consistency and reduce subjective variance, evaluators receive detailed guidelines and participate in a

calibration session before scoring. The evaluation is conducted under a double-blind setup to eliminate bias. Additionally, we compute interrater agreement using Cohen’s Kappa (Cohen, 1960) to analyze the consistency between evaluators and their alignment with the LLM-based judge.

## 4.4 Main Results

### 4.4.1 Automatic Evaluation Results

The automatic evaluation results of KEDiT, compared with those of various baseline models on the Wizard of Wikipedia and PubMed-Dialog datasets, are summarized in Table 1. KEDiT consistently outperforms all baseline models across most evaluation metrics in both datasets. In the test seen set, KEDiT achieves the highest scores, indicating superior performance in generating contextually relevant and informative responses. For the test unseen set, KEDiT demonstrates robust generalization capabilities, although it slightly lags behind SPI in F1 and BLEU. This is primarily because SPI uses predefined gold knowledge and a training method similar to LSR, which is effective but computationally expensive. However, KEDiT is designed to be applicable to a broader range of scenarios and does not rely on predefined gold knowledge. This flexibility allows KEDiT to perform well even in environments where predefined knowledge is not available.

In the PubMed-Dialog dataset, KEDiT significantly outperforms all the baselines, highlighting its exceptional ability to generate accurate and informative responses in specialized domains. Traditional knowledge-grounded models are not suitable for these scenarios because they depend on predefined knowledge, which is not available in the PubMed-Dialog dataset. The existing retrieved knowledge-augmented methods do not perform well on this dataset, likely because they are not optimized to effectively utilize the retrieved specialized knowledge. The knowledge bottleneck mechanism compresses and integrates knowledge more effectively, ensuring that KEDiT can utilize the knowledge during dialogue generation.

KEDiT also demonstrates efficiency, requiring fine-tuning of only 140M parameters, less than 2% of the LLM’s 8B parameters. These results underscore the effectiveness of KEDiT in compressing and integrating knowledge, leading to substantial improvements while maintaining efficiency.

MODEL	WoW SEEN			WoW UNSEEN			PubMed-DIALOG			TUNED
	F1	BLEU	ROUGE	F1	BLEU	ROUGE	F1	BLEU	ROUGE	PARAMS
BART <sub>large</sub>	20.54	12.10	14.73	18.38	10.53	12.78	32.30	21.17	23.53	406M
LLAMA <sub>lora</sub>	20.45	12.37	14.95	20.26	12.14	14.71	35.77	<u>23.87</u>	26.38	168M
TRANSIKG	21.31	12.77	<u>16.61</u>	19.40	11.71	15.08	–	–	–	194M
SPI-UNIFORM	<u>21.82</u>	<u>13.05</u>	16.43	<b>21.14</b>	<b>12.80</b>	<u>15.89</u>	–	–	–	141M
KAT-TSLF	20.50	12.45	15.13	19.60	11.97	14.30	<u>36.16</u>	23.52	<u>27.05</u>	198M
LLAMA <sub>rag</sub>	17.07	8.69	11.61	17.41	8.84	11.94	34.40	20.80	26.05	0M
LLAMA <sub>kpt</sub>	19.28	10.80	13.99	18.28	9.93	13.07	29.43	18.66	23.98	214M
KEDiT	<b>22.45*</b>	<b>13.87*</b>	<b>17.24*</b>	<u>21.05</u>	<u>12.63</u>	<b>15.94</b>	<b>38.63*</b>	<b>25.84*</b>	<b>28.91*</b>	140M
– KC	21.05	12.60	15.81	20.08	11.76	15.00	37.04	24.01	27.43	140M
– $\mathcal{L}_{align}$ in KC	22.01	13.16	16.92	20.64	12.03	15.81	37.93	25.28	28.40	140M
– KA-ADAPTER	18.41	9.71	12.35	17.72	9.57	12.17	33.42	20.86	25.12	3M
– KA-ATTN	22.06	13.38	16.86	20.79	12.16	15.52	38.10	25.65	28.46	138M
– KA-FFN	20.37	11.77	14.47	19.13	10.77	13.97	35.52	22.93	26.20	5M

Table 1: Automatic evaluation results on Wizard of Wikipedia (WoW) and PubMed-Dialog test sets. The best results are shown in **bold**, and the second-best results are underlined. The table also includes ablation experiments showing the performance effect of removing key components of KEDiT. Significant improvements over the best baseline are marked by \* (one-sample t-test,  $p < 0.05$ ).

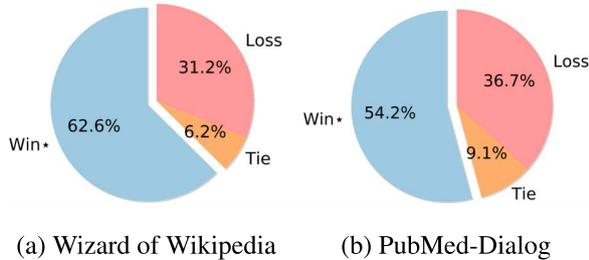


Figure 2: Pairwise comparison results of KEDiT against baseline models, showing win, tie, and loss rates. The comparisons are against SPI on the Wizard of Wikipedia test sets and Llama<sub>lora</sub> on the PubMed-Dialog test set. Significant improvements are marked with \* (binomial test,  $p < 0.01$ ).

#### 4.4.2 LLM-Based Evaluation Results

In the pairwise comparisons, we evaluate KEDiT against the best-performing baseline models in automatic evaluation metrics, namely, SPI for the Wizard of Wikipedia dataset and Llama<sub>lora</sub> for the PubMed-Dialog dataset. As shown in Figure 2, KEDiT significantly outperforms SPI on the Wizard of Wikipedia dataset, achieving a higher win rate and demonstrating a superior ability to generate relevant and informative responses. Similarly, on the PubMed-Dialog dataset, KEDiT achieves a notable win rate over Llama<sub>lora</sub>, underscoring its effectiveness in incorporating domain-specific knowledge into the dialogue generation process.

In the multi-response grading evaluation, we compare KEDiT with the best-performing models from three baseline categories on the basis

of automatic evaluation. As detailed in the upper part of Table 2, compared with these baselines, KEDiT consistently achieves higher scores for relevance, informativeness, and fluency. Interestingly, although SPI scores highly on automatic metrics, it performs worse than does Llama<sub>lora</sub> in LLM-based evaluations. This discrepancy is likely due to Llama’s robust language generation capabilities, which produce more coherent and contextually appropriate responses. SPI’s reliance on predefined knowledge may limit its adaptability, resulting in less natural dialogue flow. KEDiT, based on the Llama model, combines the strengths of both informativeness and fluency, achieving superior results in both automatic and LLM-based evaluations.

#### 4.4.3 Human Evaluation Results

The human evaluation results are summarized in the lower part of Table 2. Consistent with the LLM-based evaluation, KEDiT consistently outperforms the baseline models across all the metrics in both datasets. Notably, the improvements in relevance and informativeness scores are statistically significant in both datasets, highlighting KEDiT’s superior ability to generate responses that are not only contextually appropriate but also rich in information. The lack of significant improvement in fluency is because both the strong baseline and KEDiT already perform exceptionally well. Additionally, we observe that compared with the LLM-based evaluation, human scores

METHOD	MODEL	WIZARD OF WIKIPEDIA			PUBMED-DIALOG		
		RELEVANCE	INFORMATIVENESS	FLUENCY	RELEVANCE	INFORMATIVENESS	FLUENCY
LLM-BASED EVALUATION	LLAMA <sub>lora</sub>	3.87	2.61	4.46	4.46	3.42	4.86
	SPI-UNIFORM	3.29	2.49	4.25	–	–	–
	KAT-TSLF	3.16	2.31	4.18	3.88	3.34	4.45
	KEDiT	<b>3.92</b>	<b>2.72*</b>	<b>4.59*</b>	<b>4.85*</b>	<b>3.82*</b>	<b>4.89</b>
HUMAN EVALUATION	LLAMA <sub>lora</sub>	3.91	3.11	4.52	4.50	3.83	4.80
	SPI-UNIFORM	3.76	3.04	4.28	–	–	–
	KAT-TSLF	3.50	3.00	4.21	4.02	3.58	4.57
	KEDiT	<b>4.09*</b>	<b>3.36*</b>	<b>4.60</b>	<b>4.72*</b>	<b>4.03*</b>	<b>4.82</b>

Table 2: Multi-response grading evaluation results on Wizard of Wikipedia and PubMed-Dialog test sets. Significant improvements over the best baseline are marked by \* (independent t-test,  $p < 0.05$ ).

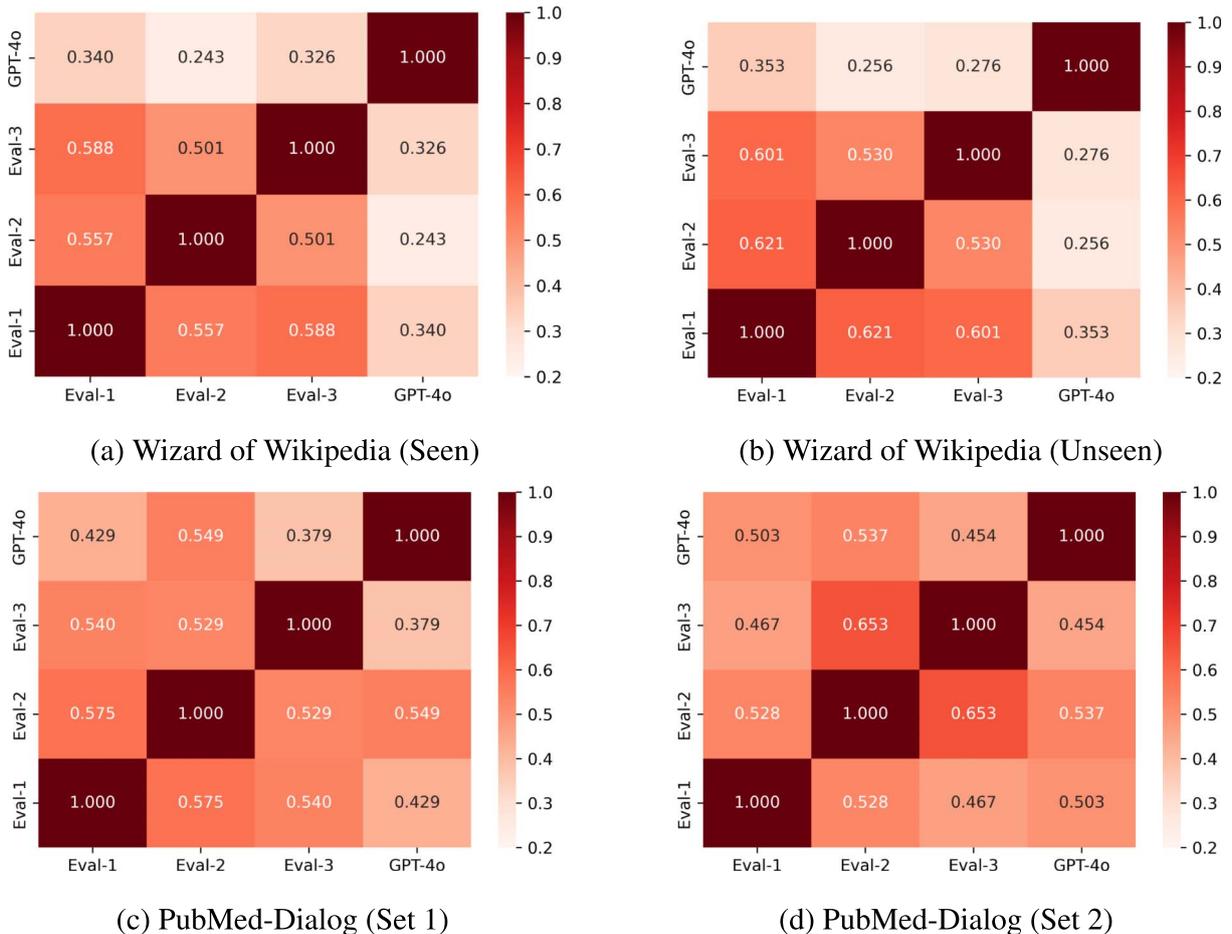


Figure 3: Heatmaps of Cohen’s Kappa coefficient matrix showing the agreement between evaluators (abbreviated as Eval-1, Eval-2, and Eval-3) and GPT-4o. Higher coefficients indicate greater agreement.

tend to be slightly more tempered, with less pronounced score differences between models for the same sample.

To quantify the agreement between human evaluators and the LLM-based judge, we compute Cohen’s Kappa coefficients, which are visualized in Figure 3. The interrater agreement between human evaluators, with Kappa coefficients between

human evaluators generally above 0.5 in most cases, indicates moderate to substantial agreement. Similarly, the agreement between human evaluators and GPT-4o also spans from fair to moderate, with Kappa coefficients generally above 0.24 and reaching 0.55 in some cases. However, lower Kappa scores are observed primarily in the Wizard of Wikipedia dataset, likely due to the

[Knowledge]	BACKGROUND Immunotherapy combined with molecular targeted therapy is increasingly popular in patients with advanced hepatocellular carcinoma (HCC)... ... CONCLUSIONS AA could reduce liver injury by preventing hepatocyte SLC7A11/GPX4 ferroptosis and improve the immunotherapy effect of anti-PD1 by boosting CD11b+CD44+PD-L1+cell population in HCC.
[Dialog Context]	User: How does immunotherapy work for treating hepatocellular carcinoma (HCC)?  Assistant: Immunotherapy helps activate the body's immune system to target and attack cancer cells in HCC. Commonly used are immune checkpoint inhibitors that target PD-1/PD-L1 proteins to enhance immune response.  User: What are the common side effects of immunotherapy in HCC patients?  Assistant: Immunotherapy can cause immune-related adverse events (irAEs) such as liver toxicity, fatigue, skin reactions, and gastrointestinal issues. Managing these side effects remains a crucial aspect of treatment.  User: Can anything be done to minimize the liver injury caused by immunotherapy?  Assistant: Research suggests that adding ascorbic acid (vitamin C) to the treatment regimen may reduce liver injury by preventing certain cellular processes that lead to hepatocyte damage. ... User: Is this approach effective in improving the overall outcomes of immunotherapy for HCC?
[Model Response]	
GPT-4o	Preliminary findings suggest that ascorbic acid not only <b>reduces liver injury</b> but also may <b>enhance the overall efficacy of immunotherapy</b> by boosting certain immune cell populations. <b>Further clinical evaluation is needed.</b>
KAT-TSLF	Yes, ascorbic acid has been shown to <b>significantly reduce liver injury</b> and <b>improve overall survival</b> in patients with HCC.
Llama <sub>lora</sub>	Yes, it has shown <b>promising results</b> by <b>enhancing the therapeutic effectiveness</b> of immunotherapy, which can lead to <b>better patient outcomes</b> .
KEDiT	The combination of immunotherapy and ascorbic acid has shown <b>promising results</b> in preclinical studies, indicating <b>potential benefits in improving treatment outcomes</b> for HCC patients. However, <b>further clinical trials are needed</b> to confirm these findings.

Figure 4: Case study from the PubMed-Dialog test set comparing KEDiT with several baselines. The dialogue shows user questions and assistant responses on immunotherapy for hepatocellular carcinoma (HCC).

shorter responses in this dataset, because GPT-4o tends to assign lower informativeness scores to brief responses. This level of agreement indicates that GPT-4o’s evaluations are reasonably consistent with human judgments, which supports the reliability of GPT-4o as a judge.

## 4.5 Analysis

### 4.5.1 Ablation Study

To assess the impact of each component in KEDiT, we perform an ablation study by removing individual modules and assessing performance on the Wizard of Wikipedia and PubMed-Dialog datasets. The lower part of Table 1 shows that removing any module results in decreased performance across automatic evaluation metrics, highlighting the importance of each component.

Specifically, removing the knowledge-aware adapter results in the most significant decrease. This is likely because, without it, the knowledge queries alone cannot effectively integrate com-

pressed knowledge into dialogue generation. The KA-Adapter provides specialized mechanisms to incorporate this knowledge, ensuring its effective utilization. Removing KA-Attn or KA-FFN also results in decreased performance, which further confirms the critical role of these submodules in facilitating knowledge integration. Removing the knowledge compression module (KC) also causes a substantial decrease, which highlights its importance in efficiently distilling essential information from retrieved knowledge. Excluding alignment loss ( $\mathcal{L}_{align}$ ) results in a smaller but noticeable degradation, which indicates the necessity of aligning compressed knowledge with the internal representations of the LLM.

### 4.5.2 Case Study

Figure 4 shows a dialogue from the PubMed-Dialog test set where the user queries about immunotherapy for hepatocellular carcinoma (HCC). When the user asks about the effectiveness of combining ascorbic acid (AA) with

MODEL	WoW SEEN	WoW UNSEEN	PMD
LLAMA <sub>lora</sub>	20.45	20.26	35.77
MISTRAL <sub>lora</sub>	20.09	19.73	35.52
QWEN <sub>lora</sub>	20.51	20.11	35.92
LLAMA <sub>redit</sub>	22.45	21.05	38.63
MISTRAL <sub>redit</sub>	21.94	20.76	37.69
QWEN <sub>redit</sub>	22.15	20.94	38.31

Table 3: F1 scores comparing LoRA-based fine-tuning and KEDiT-enhanced models across different LLMs.

immunotherapy for improving overall outcomes, KEDiT responds that this combination has shown promising results in preclinical studies, indicating potential benefits but emphasizing the need for further clinical trials. This response balances optimism with caution, providing a comprehensive and realistic assessment. In contrast, KAT-TSLF states that ascorbic acid significantly reduces liver injury and improves overall survival, which lacks a nuanced perspective on the need for further validation. Llama<sub>lora</sub> mentions the approach’s promise in enhancing therapeutic effectiveness but does not address the necessity for additional clinical trials, making its response less thorough. Compared with the response generated by GPT-4o, which highlights the potential benefits and calls for further evaluation, KEDiT similarly emphasizes the need for additional clinical trials, offering a balanced and detailed response.

#### 4.5.3 Cross-Model Generalization Analysis

To evaluate the generalizability and robustness of KEDiT across different large language models, we integrate our framework with two additional state-of-the-art open-source LLMs: Qwen2.5 (Yang et al., 2025) and Mistral-v0.3 (Jiang et al., 2023), each with 7B parameters. To ensure a fair and consistent evaluation, we conduct experiments using the same training and inference pipelines as those used for Llama-3-8B. Additionally, we fine-tune each LLM using LoRA to establish baseline performances. As shown in Table 3, KEDiT consistently outperforms LoRA-based fine-tuning across Llama-3, Qwen2.5, and Mistral-v0.3 on both the Wizard of Wikipedia and PubMed-Dialog test sets. It is worth noting that the results for Qwen<sub>redit</sub> and Mistral<sub>redit</sub> are slightly lower compared to Llama<sub>redit</sub>, primarily because Llama<sub>redit</sub> was specifically fine-tuned with hyperparameter optimization, while the same parameters were

$m$	WoW SEEN	WoW UNSEEN	PUBMED-DIALOG
2	21.32	20.58	37.06
4	21.85	20.48	37.64
8	22.01	20.97	38.37
16	<b>22.45</b>	21.05	38.63
32	22.33	<b>21.15</b>	<b>38.69</b>

Table 4: F1 scores for different numbers of knowledge queries on the Wizard of Wikipedia and PubMed-Dialog test sets.

directly applied to Qwen<sub>redit</sub> and Mistral<sub>redit</sub>. Nevertheless, the advantage of KEDiT over LoRA remains significant. Compared to LoRA, KEDiT’s knowledge compression and adapter-based integration enable more efficient and targeted utilization of external knowledge, leading to higher F1 scores in both open-domain and specialized-domain tasks. These results highlight the scalability and adaptability of KEDiT, showing that its lightweight yet powerful mechanism for knowledge integration can generalize well across different LLM families.

#### 4.5.4 Impact of Knowledge Queries

We conduct experiments to assess the effect of varying the number of knowledge queries ( $m$ ) on KEDiT’s performance, with values of  $m$  set at 2, 4, 8, 16, and 32. Our results, summarized in Table 4, reveal a positive correlation between increasing  $m$  and model performance. However, while higher  $m$  values generally improve performance, the gains beyond setting  $m$  to 16 are minimal. This diminishing return may be due to the model reaching a saturation point where additional queries contribute little new information. Moreover, setting  $m$  to 16 allows the model to be efficiently trained on a consumer-grade GPU with 24 GB of memory, whereas setting  $m$  to 32 exceeds this capacity, making training impractical on such hardware.

#### 4.5.5 Impact of Retrieval Performance

To evaluate the impact of retrieval performance on KEDiT, we conduct experiments by varying the percentage of dialogue turns containing the predefined gold knowledge sentence in the Wizard of Wikipedia dataset. Since KEDiT itself does not include a retrieval component, this indirect approach helps assess how the accuracy of retrieved knowledge affects dialogue generation. We conduct tests with percentages of 60%, 70%, 80%,

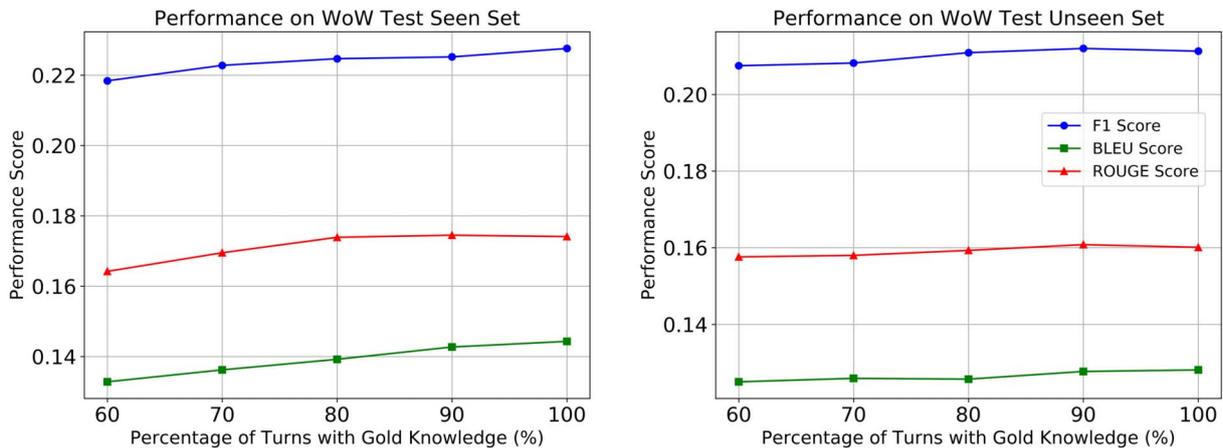


Figure 5: Performance of KEDiT with different percentages of gold knowledge retrieval.

MODEL	FACT CHECK.	ENTITY LINKING	SLOT FILLING	OPEN DOMAIN QA			
	<b>FEVER</b>	<b>AY2</b>	<b>zsRE</b>	<b>NQ</b>	<b>HOTPOTQA</b>	<b>TRIVIAQA</b>	<b>ELI5</b>
	ACCURACY			EXACT MATCH			ROUGE-L
BART	79.80	82.66	5.29	13.18	14.46	18.40	20.04
LLAMA <sub>lora</sub>	87.22	90.16	28.11	40.36	27.69	43.65	20.63
LLAMA <sub>rag</sub>	90.53	57.29	66.00	54.53	43.60	86.97	18.17
KEDiT	88.83	91.70	32.85	44.50	28.38	62.11	22.19

Table 5: Performance on knowledge intensive tasks (dev sets).

90%, and 100% and evaluate the performance of KEDiT under these conditions. As shown in Figure 5, the performance of KEDiT generally improves as the percentage of turns with gold knowledge increases. However, the overall differences are not very large, indicating that KEDiT maintains a robust level of performance even with less-than-perfect knowledge retrieval. Notably, the performance gains are more pronounced in the seen set than in the unseen set. This finding clearly suggests that the model benefits more from accurate knowledge retrieval when dealing with familiar topics. In contrast, the unseen set involves new topics, which limits the degree of improvement even with high retrieval accuracy. While the performance of KEDiT in unfamiliar domains is somewhat constrained, its lightweight nature allows for quick adaptation to new domains with minimal retraining.

#### 4.5.6 Evaluation on the KILT Benchmark

To assess the adaptability and robustness of KEDiT across diverse domains and tasks, we conduct additional experiments using the KILT benchmark (Petroni et al., 2021). This benchmark

is a widely recognized framework for evaluating models on a variety of knowledge-intensive language tasks, encompassing tasks such as fact-checking, entity linking, and open-domain question answering. We evaluate KEDiT on the following tasks included in KILT: FEVER (Thorne et al., 2018), AIDA CoNLL-YAGO (AY2; Hoffart et al., 2011), Zero Shot RE (Levy et al., 2017), Natural Questions (NQ; Kwiatkowski et al., 2019), HotpotQA (Yang et al., 2018), TriviaQA (Joshi et al., 2017), and ELI5 (Fan et al., 2019). Metrics include Accuracy, Exact Match (EM), and ROUGE-L, following KILT’s standard evaluation procedures. We leverage the off-the-shelf Contriever-MS MARCO (Izacard et al., 2022) to retrieve three relevant documents for each input.

Table 5 summarizes the results across all evaluated tasks. In evaluating Llama<sub>rag</sub>, we observe significant discrepancies between its predictions and ground-truth labels on most tasks except ELI5. When directly computing evaluation metrics, these discrepancies result in extremely low scores, often approaching zero. To address this issue and make the comparison more meaningful, we adopt a relaxed evaluation approach: If

the ground-truth label appears in the predictions of Llama<sub>rag</sub>, the prediction is considered correct. This adjustment has significantly inflated metric values for Llama<sub>rag</sub>. Despite these adjustments for Llama<sub>rag</sub>, KEDiT consistently outperforms baselines, demonstrating superior adaptability and robustness across knowledge-intensive tasks. The ELI5 task stands out in the evaluation as it features long-form answers, which align well with KEDiT’s strengths. This suggests that KEDiT’s design makes it particularly well-suited for tasks requiring detailed and extended responses, a characteristic often observed in dialogue scenarios.

## 5 Discussion

### 5.1 Potential Biases in PubMed-Dialog

While the PubMed-Dialog dataset is constructed to enhance knowledge-grounded dialogue in the biomedical domain, it may inherit biases from its data sources. First, PubMed primarily consists of peer-reviewed research articles, which may introduce selection bias by over-representing academic perspectives while underrepresenting clinical or patient viewpoints. Second, since the dialogues are synthesized using GPT-4o, there is potential for stylistic and terminological biases that could affect model generalization to non-expert users. Moreover, models from the GPT family (or those pretrained on GPT-generated data) may exhibit inflated performance due to inherent similarities in language patterns and knowledge representation. This could lead to an unfair advantage for such models compared to others not exposed to GPT-generated content during training. Finally, the dataset captures knowledge at a fixed point in time, meaning that newer medical discoveries may not be adequately reflected. While these biases are inevitable to some extent, we employ iterative validation processes to minimize their impact. Additionally, we plan to explore alternative evaluation strategies in future work, such as cross-utilizing different LLMs for both generation and evaluation, to further mitigate potential overfitting to GPT-generated text.

### 5.2 Limitations of Automatic Metrics in Dialogue Evaluation

While BLEU and ROUGE remain widely used for the automatic evaluation of text generation models, they have notable limitations in dialogue evaluation. These metrics primarily focus on

$n$ -gram overlaps with reference responses, making them insufficient for assessing contextual coherence, factual correctness, and diversity of generated responses. Prior studies (Liu et al., 2016; Novikova et al., 2017) have demonstrated that such surface-based metrics correlate poorly with human judgments in dialogue settings. Given these shortcomings, we complement our evaluation with LLM-based scoring and human assessments to better capture the conversational quality and informativeness of the generated responses.

## 6 Conclusion

This paper presents KEDiT, a novel approach for improving knowledge-grounded dialogue generation in LLMs. KEDiT effectively addresses the limitations of LLMs in utilizing up-to-date and domain-specific knowledge by compressing external knowledge into learnable parameters and integrating them using a lightweight adapter. To support this evaluation, we create the PubMed-Dialog dataset, which provides a benchmark for assessing the ability of the model to address specialized biomedical knowledge. Our extensive experiments on the Wizard of Wikipedia and PubMed-Dialog datasets demonstrate that, compared with existing methods, KEDiT significantly improves the contextual relevance and informativeness of generated responses. Further analysis highlights KEDiT’s robust performance even with varying retrieval accuracy, maintaining high levels of contextual relevance and informativeness. Although KEDiT shows slightly reduced performance on unseen domains, its design allows for efficient retraining and deployment in environments requiring frequent updates to knowledge. Future work will focus on enabling dynamic knowledge updates, enhancing generalization to unseen domains, and integrating multimodal knowledge for greater adaptability and practicality.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China (no. 62376051) and the Anhui Provincial Natural Science Foundation (2408085QF188). We thank the anonymous ACL reviewers and the action editor, Xiaojun Wan, for their insightful feedback. We also appreciate the twelve evaluators who contributed to our human evaluation.

## References

- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.
- Jiaqi Bai, Zhao Yan, Ze Yang, Jian Yang, Xinnian Liang, Hongcheng Guo, and Zhoujun Li. 2023. Knowprefix-tuning: A two-stage prefix-tuning framework for knowledge-grounded dialogue generation. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, pages 525–542, Cham. Springer Nature Switzerland. [https://doi.org/10.1007/978-3-031-43415-0\\_31](https://doi.org/10.1007/978-3-031-43415-0_31)
- David Barber and Felix Agakov. 2003. The IM algorithm: A variational approach to information maximization. In *Advances in Neural Information Processing Systems*, volume 16. MIT Press.
- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46. <https://doi.org/10.1177/001316446002000104>
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

- Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1346>
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.
- Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.
- Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.
- Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth ee Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825v1.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1259>
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics. <https://doi.org/10.18653/v1/P17-1147>
- Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International*

- Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.
- Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. 2021. Compacter: Efficient low-rank hypercomplex adapter layers. In *Advances in Neural Information Processing Systems*, volume 34, pages 1022–1035, Curran Associates, Inc.
- Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. In *International Conference on Learning Representations*.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466. [https://doi.org/10.1162/tacl\\_a\\_00276](https://doi.org/10.1162/tacl_a_00276)
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.emnlp-main.243>
- Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics. <https://doi.org/10.18653/v1/K17-1034>
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.703>
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.
- Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.353>
- Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2024. RA-DIT: Retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations*.
- Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.
- Shilei Liu, Xiaofeng Zhao, Bochao Li, Feiliang Ren, Longhui Zhang, and Shujuan Yin. 2021.

- A three-stage learning framework for low-resource knowledge-grounded dialogue generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2262–2272, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.emnlp-main.173>
- Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-short.8>
- Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2024. GPT understands, too. *AI Open*, 5:208–215. <https://doi.org/10.1016/j.aiopen.2023.08.012>
- Chuan Meng, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tengxiao Xi, and Maarten de Rijke. 2021. Initiative-aware self-supervised learning for knowledge-grounded conversations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21*, pages 522–532, New York, NY, USA. Association for Computing Machinery. <https://doi.org/10.1145/3404835.3462824>
- Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics. <https://doi.org/10.18653/v1/D17-1238>
- OpenAI. 2022. Introducing ChatGPT.
- OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774v3.
- Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: A benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.naacl-main.200>
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.
- Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331. <https://doi.org/10.1162/tacl.a.00605>
- Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*.
- Siqi Shen, Veronica Perez-Rosas, Charles Welch, Soujanya Poria, and Rada Mihalcea. 2022. Knowledge enhanced reflection generation for counseling dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3096–3107, Dublin, Ireland. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.221>
- Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. REPLUG: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384, Mexico

- City, Mexico. Association for Computational Linguistics.
- Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17. [https://doi.org/10.1162/tacl\\_a.00530](https://doi.org/10.1162/tacl_a.00530)
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics. <https://doi.org/10.18653/v1/N18-1074>
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMa: Open and efficient foundation language models. *CoRR*, abs/2302.13971v1.
- Yan Xu, Etsuko Ishii, Samuel Cahyawijaya, Zihan Liu, Genta Indra Winata, Andrea Madotto, Dan Su, and Pascale Fung. 2022. Retrieval-free knowledge-grounded dialogue response generation with adapters. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 93–107, Dublin, Ireland. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.dialdoc-1.10>
- Yan Xu, Deqian Kong, Dehong Xu, Ziwei Ji, Bo Pang, Pascale Fung, and Ying Nian Wu. 2023. Diverse and faithful knowledge-grounded dialogue generation via sequential posterior inference. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 38518–38534. PMLR.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *CoRR*, abs/2412.15115v2.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics. <https://doi.org/10.18653/v1/D18-1259>
- Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations*.
- Bo Zhang, Hui Ma, Jian Ding, Jian Wang, Bo Xu, and Hongfei Lin. 2025. Distilling implicit multimodal knowledge into large language models for zero-resource dialogue generation. *Information Fusion*, 118:102985. <https://doi.org/10.1016/j.inffus.2025.102985>
- Bo Zhang, Jian Wang, Hongfei Lin, Hui Ma, and Bo Xu. 2022. Exploiting pairwise mutual information for knowledge-grounded dialogue. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2231–2240. <https://doi.org/10.1109/TASLP.2022.3161151>
- Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023a. Adaptive budget allocation for parameter-efficient fine-tuning.

In *The Eleventh International Conference on Learning Representations*.

Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. 2024. LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023b. How do large language models capture the ever-changing world knowledge? A review of recent advances. In *Proceedings of the 2023*

*Conference on Empirical Methods in Natural Language Processing*, pages 8289–8311, Singapore. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.516>

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

## A KEDiT Training and Inference

Algorithms 1 and 2 provide a step-by-step breakdown of our training and inference processes.

---

### Algorithm 1 KEDiT Training

---

- 1: **Input:** Knowledge dataset  $\mathcal{D}_k$ , Knowledge-grounded dialogue dataset  $\mathcal{D}_d$ , Pre-trained BERT, Pre-trained LLM
  - Phase 1: Knowledge Compression Training**
  - 2: Initialize knowledge bottleneck with Q-Former and frozen BERT ▷ Denoted as  $\phi$
  - 3: Freeze parameters of the pre-trained LLM ▷ Denoted as  $\psi$
  - 4: **for** each  $K \in \mathcal{D}_k$  **do**
  - 5:   Encode  $K$  using the knowledge bottleneck to obtain  $Z$  based on Eq. (2) ▷  $p_\phi(Z|K)$
  - 6:   Reconstruct  $K$  from  $Z$  using the LLM to obtain  $\hat{K}$  ▷  $q_\psi(K|Z)$
  - 7:   Reconstruct  $Z$  from  $K$  using the LLM to obtain  $\hat{Z}$  ▷  $q_\psi(Z|K)$
  - 8:   Compute  $\mathcal{L}_{\text{recon}}$  and  $\mathcal{L}_{\text{align}}$  as defined in Eq. (4) and Eq. (5)
  - 9:   Update  $\phi$  by minimizing  $\mathcal{L}_{\text{kc}}$  as defined in Eq. (6)
  - 10: **end for**
  - Phase 2: Knowledge Integration Training**
  - 11: Integrate KA-Adapter with the frozen LLM ▷ Denoted as  $\theta$
  - 12: Freeze  $\phi$  except for  $\mathbf{W}_z$
  - 13: **for** each  $(C, K, R) \in \mathcal{D}_d$  **do**
  - 14:   Encode  $K$  using the tuned knowledge bottleneck to obtain  $Z$  ▷  $p_\phi(Z|K)$
  - 15:   Concatenate  $C$  and  $Z$  to form the input
  - 16:   Compute the likelihood of the response  $R$  based on Eq. (7), Eq. (8) and Eq. (9) ▷  $p_\theta(R|C, Z)$
  - 17:   Update  $\theta$  and  $\mathbf{W}_z$  by minimizing  $\mathcal{L}_{\text{gen}}$  as defined in Eq. (10)
  - 18: **end for**
  - 19: **Output:** Trained knowledge bottleneck  $\phi$  and LLM integrated with KA-Adapter  $\theta$
- 

---

### Algorithm 2 KEDiT Inference

---

- 1: **Input:** Dialogue context  $C$ , Retrieved knowledge  $K$ , Trained knowledge bottleneck  $\phi$ , Trained LLM with KA-Adapter  $\theta$
  - 2: Encode  $K$  using the knowledge bottleneck module to obtain  $Z$  ▷  $p_\phi(Z|K)$
  - 3: Initialize response  $R_0 \leftarrow \emptyset$ ,  $t \leftarrow 1$
  - 4: **while** not end-of-sequence **do**
  - 5:   Concatenate  $C$ ,  $Z$  and  $R_{<t}$  to form the input
  - 6:   Compute the probability of the next token  $R_t$  based on Eq. (7), Eq. (8) and Eq. (9) ▷  $p_\theta(R_t|C, Z, R_{<t})$
  - 7:   Sample or select  $R_t$  based on  $p_\theta(R_t|R_{<t}, C, Z)$
  - 8:   Append  $R_t$  to the response:  $R \leftarrow R \oplus R_t$
  - 9:   Update  $t \leftarrow t + 1$
  - 10: **end while**
  - 11: **Output:** Generated response  $R$
- 

## B Prompt Design and Iterative Validation Process for the PubMed-Dialog Dataset

**Iterative Validation Process** In this work, we employ an iterative validation process to ensure the quality and faithfulness of the PubMed-Dialog dataset. This process involves the following steps:

1. **First Round (Initial Evaluation):** Each dialogue is evaluated on three key criteria: *Source Consistency*, *Internal Consistency*, and *Factual Accuracy*. Scores ranging from 1 to 5 are assigned for each criterion, where 1 indicates poor quality and 5 represents high quality.
  - *Source Consistency:* Does the dialogue accurately and faithfully represent the information from the abstract?
  - *Internal Consistency:* Is the dialogue coherent and logically consistent within itself?

- **Factual Accuracy:** Does the dialogue contain accurate medical information that is consistent with established knowledge?
2. **Regeneration:** Dialogues that score below a threshold (i.e., scores < 5) are flagged for regeneration. These flagged dialogues are revised using GPT-4o, with a focus on eliminating hallucinations and ensuring a closer alignment with the source content, while adhering to the evaluation criteria.
  3. **Re-evaluation and Iterative Refinement:** The regenerated dialogues are evaluated again using the same criteria. If they still do not meet the required standards, they are flagged for further regeneration and re-evaluated until they meet the desired quality threshold.

Table 6 summarizes the results of the three rounds of evaluation and regeneration for the PubMed-Dialog dataset. The iterative process has significantly improved dialogue quality across all criteria, particularly addressing initial deficiencies in *Source Consistency*. While *Internal Consistency* required minimal corrections and *Factual Accuracy* improved substantially, minor alignment issues with biomedical knowledge persist, reflecting the complexity of the domain. Although the dataset does not achieve absolute perfection, it provides a robust benchmark for knowledge-grounded dialogue generation in specialized domains.

ROUND	SOURCE CONSISTENCY ( $\leq 3/4/5$ )	INTERNAL CONSISTENCY ( $\leq 3/4/5$ )	FACTUAL ACCURACY ( $\leq 3/4/5$ )	TOTAL DIALOGUES EVALUATED	PERCENTAGE OF HIGH-QUALITY DIALOGUES
Round 1	57 / 1,961 / 8,912	1 / 40 / 10,889	19 / 686 / 10,225	10,930	8,884 (81.28%)
Round 2	1 / 712 / 1,333	0 / 3 / 2,043	0 / 148 / 1,898	2,046	1,314 (64.22%)
Round 3	0 / 399 / 333	0 / 3 / 729	0 / 70 / 662	732	321 (43.85%)

Table 6: Results of the iterative evaluation and regeneration process, showing the number of dialogues scoring  $\leq 3$ , 4, and 5 in each of the three evaluation criteria across all three validation rounds.

**Prompt Design** Figure 6 shows the system prompt used to generate the PubMed-Dialog dataset.

**[System]**  
 You are an AI assistant specialized in medical topics.  
 You (Assistant) have access to a detailed text-based medical document, such as a research paper or clinical guideline, which is not visible to the person (User). The document contains specific medical knowledge, terms, and data relevant to a wide range of medical inquiries.  
 Your task is to generate a conversation between the person (User) asking about a medical topic and you (Assistant) responding based on the knowledge from the document. The conversation should proceed as though both the User and Assistant are discussing the topic openly, without directly referring to the document itself.  
 Below are requirements for generating the questions and answers in the conversation:

- Avoid directly quoting or referring to specific facts, terms, abbreviations, dates, numbers, or names from the document, as these may reveal the conversation is based on the textual information, rather than a general discussion.
- Do not use phrases like "mentioned in the document", "according to the text" or "the study". Instead, present the information as general knowledge.
- Ensure questions are diverse and cover a range of aspects related to the medical topic being discussed.
- The conversation should include at least 2-3 turns of questions and answers about the medical topic.
- Ensure that there is at most one longer answer, and that other answers should be shorter and more direct to maintain the natural flow of the conversation (less than 30 word).
- Answer responsibly, avoiding overconfidence, and do not provide medical advice or diagnostic information. Encourage the user to consult a healthcare professional for personalized advice.

Figure 6: Example of the system prompt for generating PubMed-Dialog datasets.

## C Prompt Templates for LLM-Based Evaluation

Figures 7 and 8 show the prompts used for pairwise comparison and multi-response grading, respectively.

**[System]**

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. You should choose the assistant that follows the user’s instructions and answers the user’s question better. Your evaluation should consider factors such as the relevance, informativeness, accuracy, and coherence of their responses. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.

---

**[Example input]**

[Dialogue History]  
{history}

[The Start of Assistant A’s Answer]  
{answer1}

[The End of Assistant A’s Answer]

[The Start of Assistant B’s Answer]  
{answer2}

[The End of Assistant B’s Answer]

Figure 7: Example of the prompt used for pairwise comparison.

**[System]**

Please act as an impartial judge and evaluate the quality of the response provided by several AI assistants to the user question displayed below. Your evaluation should consider three factors: **relevance**, **informativeness**, and **fluency** of the response. Begin your evaluation by providing a short explanation for each response, highlighting the strengths and weaknesses in relation to the three factors. Be as objective as possible. After providing your explanation, please rate each response on a scale of 1 to 5 for each factor by strictly following this format: "[[{{model}}\_{{factor}}\_{{rating}}]]", for example: "Rating: [[A\_relevance\_5]]", "Rating: [[B\_informativeness\_3]]", and so on.

---

**[Example input]**

[Dialogue History]  
{history}

[The Start of Assistant A’s Answer]  
{answer1}

[The End of Assistant A’s Answer]

[The Start of Assistant B’s Answer]  
{answer2}

[The End of Assistant B’s Answer]

[The Start of Assistant C’s Answer]  
{answer3}

[The End of Assistant C’s Answer]

Figure 8: Example of the prompt used for multi-response grading.

## D Instruction Templates for KEDiT Training and Inference

Table 7 presents the instruction templates used for KEDiT training and inference, and Table 8 provides corresponding example prompts.

Task	Instruction Template
$q_\psi(Z K)$	$\langle  start\_header\_id  \rangle system \langle  end\_header\_id  \rangle \backslash n \backslash n$ $\{system\_prompt\} \{knowledge\} \langle  eot\_id  \rangle$ $\langle  start\_header\_id  \rangle user \langle  end\_header\_id  \rangle \backslash n \backslash n$ $\{k2z\_prompt\} \langle  eot\_id  \rangle$ $\langle  start\_header\_id  \rangle assistant \langle  end\_header\_id  \rangle \backslash n \backslash n$ $\langle Z_1 \rangle \dots \langle Z_m \rangle \langle  eot\_id  \rangle$
$q_\psi(K Z)$	$\langle  start\_header\_id  \rangle system \langle  end\_header\_id  \rangle \backslash n \backslash n$ $\{system\_prompt\} \langle Z_1 \rangle \dots \langle Z_m \rangle \langle  eot\_id  \rangle$ $\langle  start\_header\_id  \rangle user \langle  end\_header\_id  \rangle \backslash n \backslash n$ $\{z2k\_prompt\} \langle  eot\_id  \rangle$ $\langle  start\_header\_id  \rangle assistant \langle  end\_header\_id  \rangle \backslash n \backslash n$ $\{knowledge\} \langle  eot\_id  \rangle$
$p_\theta(R C, Z)$	$\langle  start\_header\_id  \rangle system \langle  end\_header\_id  \rangle \backslash n \backslash n$ $\{system\_prompt\} \langle Z_1 \rangle \dots \langle Z_m \rangle \langle  eot\_id  \rangle$ $\langle  start\_header\_id  \rangle user \langle  end\_header\_id  \rangle \backslash n \backslash n$ $\{utterance_1\} \langle  eot\_id  \rangle$ $\langle  start\_header\_id  \rangle assistant \langle  end\_header\_id  \rangle \backslash n \backslash n$ $\{utterance_2\} \langle  eot\_id  \rangle$ $\dots$ $\langle  start\_header\_id  \rangle user \langle  end\_header\_id  \rangle \backslash n \backslash n$ $\{utterance_{l-1}\} \langle  eot\_id  \rangle$ $\langle  start\_header\_id  \rangle assistant \langle  end\_header\_id  \rangle \backslash n \backslash n$ $\{utterance_l\} \langle  eot\_id  \rangle$

Table 7: Instruction template used for KEDiT training and inference.  $\langle Z_1 \rangle \dots \langle Z_m \rangle$  are special markers denoting compressed knowledge vectors  $Z$ .  $\{knowledge\}$  represents raw retrieved knowledge input from  $\mathcal{D}_k$ , and  $\{utterance\}$  represents dialogue utterances between the user and assistant from  $\mathcal{D}_d$ .

Prompt Type	Prompt Examples
$\{system\_prompt\}$	You are a knowledge-based assistant. Use the following knowledge context to answer questions or engage in conversation. \n Knowledge:
$\{k2z\_prompt\}$	Identify and list the key information present in the detailed text. Extract the core pieces of key information that summarize the knowledge provided. What are the main themes or key pieces of information depicted in the text? List them. Summarize the text into essential pieces of key information. Distill the primary pieces of information from the text into concise descriptors. From the detailed knowledge described, what are the central pieces of key information? Determine the main pieces of key information that capture the essence of the text provided. What key pieces of information would you use to index the information described here?
$\{z2k\_prompt\}$	Describe the knowledge context. Provide a detailed description of the knowledge context. Can you explain what the knowledge context consisted of? Thoroughly outline the details of the knowledge context used. Provide a comprehensive overview of the knowledge used in the context. Elaborate on the content of the knowledge context used. What information did the knowledge context contain? Please describe in detail. Provide an in-depth explanation of the content covered in the knowledge context.

Table 8: Examples of prompts corresponding to the instruction templates in Table 7.

## E Additional Experiments

### E.1 Impact of Knowledge Encoder

We evaluate the impact of the knowledge encoder by replacing BERT with DeBERTaV3 (He et al., 2023) in the knowledge bottleneck module. Table 9 presents the results, showing that DeBERTa yields slight improvements on some metrics. However, these gains are minimal, indicating that the encoder choice has limited influence on overall performance.

MODEL	WoW SEEN			WoW UNSEEN			PUBMED-DIALOG		
	F1	BLEU	ROUGE	F1	BLEU	ROUGE	F1	BLEU	ROUGE
KEDiT <sub>bert</sub>	22.45	13.87	17.24	<b>21.05</b>	12.63	<b>15.94</b>	38.63	25.84	<b>28.91</b>
KEDiT <sub>deberta</sub>	<b>22.48</b>	<b>13.99</b>	<b>17.27</b>	20.97	<b>12.85</b>	15.91	<b>38.66</b>	<b>25.97</b>	28.89

Table 9: Performance comparison of KEDiT with BERT (KEDiT<sub>bert</sub>) and DeBERTa (KEDiT<sub>deberta</sub>) as knowledge encoders on WoW and PubMed-Dialog test sets.