LLM-Enhanced Query Generation and Retrieval Preservation for Task-Oriented Dialogue

Jiale Chen[1], Xuelian Dong[1], Wenxiu Xie[2], Ru Peng[3], Kun Zeng[4], Tianyong Hao[1]*,
[1]School of Computer Science, South China Normal University, China
[2]School of Computer Science, Guangdong Polytechnic Normal University, China
[3]School of Computer Science, Zhejiang University, China
[4]School of Computer Science and Engineering, Sun Yat-sen University, China
{jlchen, xldong}@m.scnu.edu.cn, wxxie@gpnu.edu.cn, rupeng@zju.edu.cn,
zengkun2@mail.sysu.edu.cn, haoty@m.scnu.edu.cn

Abstract

Knowledge retrieval and response generation are fundamental to task-oriented dialogue systems. However, dialogue context frequently contains noisy or irrelevant information, leading to sub-optimal result in knowledge retrieval. One possible approach to retrieving knowledge is to manually annotate standard queries for each dialogue. Yet, this approach is hindered by the challenge of data scarcity, as human annotation is costly. To solve the challenge, we propose an LLM-enhanced model of query-guided knowledge retrieval for task-oriented dialogue. It generates high-quality queries for knowledge retrieval in task-oriented dialogue solely using low-resource annotated queries. To strengthen the performance correlation between response generation and knowledge retrieval, we propose a retrieval preservation mechanism by further selecting the most relevant knowledge from retrieved top-K records and explicitly incorporating these as prompts to guide a generator in response generation. Experiments on three standard benchmarks demonstrate that our model and mechanism outperform previous state-of-the-art by 3.26% on average with two widely used evaluation metrics.

1 Introduction

Task-oriented dialogue (TOD) systems are designed to fulfill user demands in a human-computer dialogue manner (Chen et al., 2017), including tasks like restaurant reservations, hotel recommendations, etc. Unlike open-domain dialogue systems such as ChatGPT, TOD systems are required to generate informative responses to users in a minimum number of dialogue turns (Ni et al., 2023). Typically, these systems rely on an external Knowledge Base (KB) to retrieve necessary records before response generation. Each knowledge record consists of multiple attributes, such as "*address*" and "*phone*". Some methods attempt to encode the knowledge records into a memory module and to point out the relevant knowledge records for response generation based on dialogue context (Wu et al., 2019; Qin et al., 2020). Recently, Pre-trained Language Models (PLM) have been utilised to assist in response generation by taking the entire linearized knowledge base as input (Rony et al., 2022; Wang et al., 2022; Dong and Chen, 2023; Chen et al., 2024). However, the input knowledge sequence can easily become too long to be fed into the PLM-based model, as KBs typically contain thousands or even millions of records.

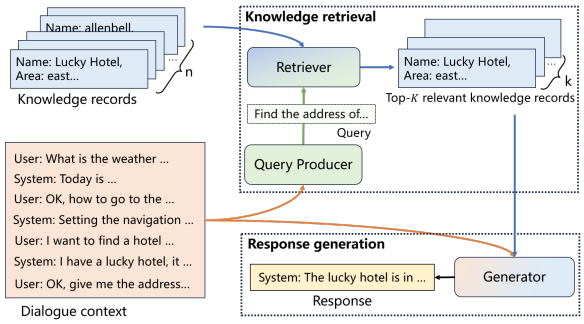


Figure 1: Illustration of the TOD system with independent retriever and query producer.

Consequently, the scalability of knowledge retrieval poses a key challenge for TOD. Addressing this challenge, an independently scalable retriever is proposed by Su et al. (2022) to retrieve the most relevant knowledge records from KB recently . It computes the similarity between a dialogue context and knowledge records to retrieve the top-K relevant records (Shen et al., 2023; Shi et al., 2023; Chen et al., 2025). Nevertheless, the dialogue context frequently contains misleading information, such as noisy, irrelevant, or outdated content. This can lead to low precision or even erroneous knowledge retrieval. For example, one user might ask for a restaurant reservation after several restaurant recommendations. The previously recommended restaurants may confuse the knowledge retrieval

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2025, pages 14307–14321

process. To alleviate the misleading information involved in knowledge retrieval, Tian et al. (2022) propose a query producer to generate a query for knowledge retrieval based on dialogue context, as shown in Figure 1, and train it via annotated query labels.

Annotating the query manually for each dialogue is a time-consuming and labor-intensive task. Thus, the current task-oriented dialogue also faces the issue of data scarcity. A potential solution is generating pseudo queries for unlabeled dialogues. However, efficiently generating high-quality pseudo queries for unlabeled dialogues to construct training pairs in order to train a query producer remains an obstacle. Furthermore, recent works identify a weak performance correlation between knowledge retrieval and response generation (Shen et al., 2023; Wan et al., 2023; Shi et al., 2023), which suggests that enhancing the precision of knowledge retrieval may weakly strengthen the response generator ability to generate more appropriate and accurate responses. This poses a difficulty in developing a TOD system to generate a high knowledge precision response.

To address these problems, we propose an LLM-enhanced model of Query-guided knowledge retrieval in TOD system (LQ-TOD). To the best of our knowledge, LQ-TOD is the first TOD system that focus the issue of query data scarcity. To alleviate the data scarcity of annotated queries in knowledge retrieval, we employ few-shot learning to generate a pseudo query. To prevent significant bias in the single pseudo query generated by an LLM, we generate a candidate pseudo query set for a dialogue context using multiple LLMs to avoid bias. Subsequently, a pseudo query selection method is introduced to select a high-quality pseudo query from the candidate pseudo query set. The selected high-quality pseudo query and its corresponding dialogue context are utilised to construct a training pair for further training the query producer. Moreover, a Retrieval Preservation Mechanism (RPM) is proposed to further enhance the performance correlation between response generation and knowledge retrieval. This RPM selectively retains the most relevant knowledge records with relevant attributes as prompts based on current user utterance. Finally, the prompts are explicitly incorporated to guide the generator in the response generation phase. In summary, this paper presents three contributions:

- A novel model of query-guided knowledge retrieval for task-oriented dialogue named LQ-TOD is proposed. To alleviate data scarcity of annotated queries, the LQ-TOD leverages few-shot learning to generate candidate pseudo queries and a scorer to select high-quality queries to further train the query producer.
- We propose a retrieval preservation mechanism RPM, which is designed to select the most relevant knowledge as prompts from retrieved top-K knowledge records to guide the generator in response generation. The RPM strengthens performance correlation between response generation and knowledge retrieval.
- Experiments on three publicly available datasets demonstrate our proposed model and mechanism outperform state-of-the-art (SOTA) baselines. Furthermore, a series of experiments validates the effectiveness of both the LQ-TOD and the RPM for query-guided knowledge retrieval and response generation.

2 Related Work

2.1 Task-Oriented Dialogue

Traditional end-to-end TOD systems employ different strategies to incorporate KB. First, the KB is embedded into a memory network or is simply embedded as parameters of the model. Madotto et al. (2018) embed the knowledge and dialogue context to a memory network for response generation, while Wu et al. (2019) introduce a global to local pointer for memory network to point out the knowledge needed to be used in response. Qin et al. (2020) design a shared encoder to embed multi-domain knowledge. Additionally, Wu et al. (2022) use a graph attention network to learn the intrinsic information in the dialogue context and knowledge graph. Madotto et al. (2020) embed the KB into model parameters by supervise training manner. This type of strategy necessitates loading all KBs into the model, posing challenges for updates when the KB changes dynamically.

Second, PLMs are utilised to encode an entire linearized knowledge base. DialoKG (Rony et al., 2022) and PluDG (Dong and Chen, 2023) directly integrate linearized knowledge and dialogue context as PLM input, incorporating relevant knowledge during response generation. Inspired by prefix tree, Ding et al. (2024) design a "prefix trie" constructed from KBs, unifying knowledge retrieval and response generation into a generation task.

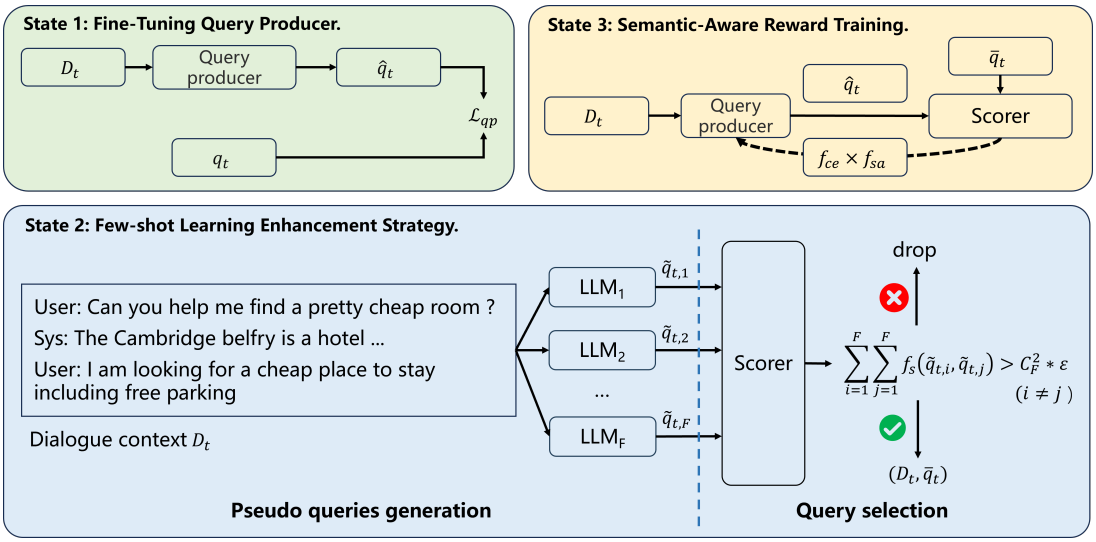


Figure 2: Illustration of the query producer within the LQ-TOD training process, which is divided into three stages.

However, this strategy relies on the PLMs, the context length of the PLM imposing a limitation on the number of KBs that can be processed.

Third, an independent knowledge retriever is adopted. Shi et al. (2023) propose a retrieval-generation architecture using the feedback from the generator as positive and negative feedback to train the retriever. MAKER (Wan et al., 2023) decouples knowledge retrieval from response generation and employs a multi-grained knowledge retrieval mechanism to fetch a set of records from an external knowledge base. Existing works just use recent K-turns dialogue context as a query to retrieve knowledge from KB, which leads to sub-optimal knowledge records retrieved for there is some noisy information in recent K-turns dialogue.

2.2 Search Query Generation

LLMs significantly improved the performance of various downstream artificial intelligence tasks (Ge et al., 2024). However, LLMs have an issue of generating misinformation, which is known as hallucination (Ji et al., 2023). To address this issue, researchers have investigated how to integrate external knowledge search engines with LLMs (Qi et al., 2019; Wang et al., 2023). Central to these external knowledge search engines is the query producer, which is utilised to formulate search queries that retrieve relevant knowledge. Wang et al. (2023) propose a dialogue model to dynamically access information from search engines and generate responses. SemiDQG (Huang et al., 2024) is a semi-supervised dialogue query generation framework that trains a response feedback module to enhance the search query producer. In contrast, Wang et al. (2023) use prompt learning to generate search queries instead of training an independent query producer. However, the efficacy of prompt-based learning is heavily contingent upon the capabilities of LLMs, exhibiting significant variability in query generation across different LLMs when provided with the same instructional prompt.

Inspired by search query generation, we propose a few-shot learning enhancement strategy to address data scarcity of annotated queries. In addition, to enhance the performance correlation between knowledge retrieval and response generation, we introduced a retrieval preservation mechanism.

3 The Method

3.1 Problem Definition

Assuming there is a dialogue context of t turns, it is denoted as $D_t = \{u_1, r_1, u_2, ..., u_t\}$. The u_t and r_t represent the user utterance and system response in the t-th turn. For each D_t in labeled data, there is an annotated query q_t used for knowledge retrieval. In contrast, for each D_t in unlabeled data, we need to construct a pseudo query $\bar{q}_t$ to construct a training pair $(D_t, \bar{q}_t)$ for the query producer. The knowledge base $\mathcal{E} = \{I_1, I_2, I_3, ..., I_N\}$ consisted of N knowledge records. Each knowledge record I_n is consisted of M attribute-value pairs, $I_n = \{(a^1, v^1), (a^2, v^2), ..., (a^M, v^M)\}$. The TOD system generates the t-th turn system response r_t, based on the dialogue context D_t and the knowledge base $\mathcal{E}$.

3.2 The Query Producer

The training process of our proposed query producer can be divided into three stages, as shown in Figure 2. In the first stage, given the dialogue context D_t and corresponding annotated query q_t, we train the query producer on the labeled training pair (D_t, q_t). In the second stage, a few-shot learning enhancement strategy is designed. It uses multiple LLMs to generate candidate pseudo queries $\widetilde{q}_i$ to form a candidate pseudo query set $\widetilde{Q}_t$ for the same dialogue context D_t. Next, a scorer is designed to compute the similarity among the generated pseudo queries from the set $\widetilde{Q}_t$. When all similarity scores of a pseudo query exceeds a threshold $C_F^2 * \varepsilon$ (see Section 3.2.2 for details), the pseudo query is used to construct a high-quality pseudo pair $(D_t,\ \bar{q}_t)$. In the third stage, we use semantic-aware reward training to further improve the query producer using high-quality pseudo pairs, and a scorer provides rewards as fine-grained training signals.

3.2.1 Fine-Tuning Query Producer

The query producer is designed to generate a query based on dialogue context to retrieve knowledge. First, we train the query producer on labeled data through supervised learning. Formally, for the t-th dialogue context D_t, a seq2seq architecture language model (LM) (Raffel et al., 2020) is utilised to generate the corresponding query $\hat{q}_t$, defining as follows:

$$\hat{q}_t = \mathrm{LM}\left(D_t\right) \tag{1}$$

Subsequently, a Cross-Entropy (CE) loss is used to train the query producer as Equation 2.

$$\mathcal{L}_{qp} = -\log p\left(q_t | D_t, \theta_{qp}\right) \tag{2}$$

θ_{qp} are parameters of the query producer.

3.2.2 Few-shot Learning Enhancement Strategy

At the same time, we input the labeled data as examples to the LLMs. A prompt[1] is designed to construct examples resampled from labeled data, denoted as p^{fsl}. Utilizing the few-shot learning of LLM, a pseudo query is generated. To avoid errors caused by a single LLM, for each dialogue context D_t, there are multiple LLMs to generate candidate pseudo queries $\widetilde{q}_{t,i}$. Formally, the candidate pseudo query generation is expressed as Equation 3.

$$\widetilde{q}_{t,i} = \mathrm{LLM}_i\left(p^{fsl}; D_t\right),\ i \in F \tag{3}$$

[1]More design is shown in Appendix A in detail.

Specifically, the number of LLMs used here is defined as F ($F = 3$). The $\widetilde{q}_{t,i}$ is the candidate pseudo query generated from i-th LLM according to the prompt and an unlabeled dialogue context D_t. To measure the quality of the generated pseudo queries, a scorer f_s is used to calculate the similarity among the candidate pseudo queries. Only those candidate pseudo queries whose summed-up similarity score $\bar{s}_t$ is higher than the threshold $C_F^2 * \varepsilon$ (C_F^2 denotes combinations, $\varepsilon = 0.5$) are used as the final pseudo query $\bar{q}_t$ ($\bar{q}_t = \{\widetilde{q}_{t,i}\}_{i=1}^{F}$). In contrast, if $\bar{s}_t$ is lower than ε, the dialogue context D_t is excluded from further training of the query producer. The calculation is expressed as Equation 4.

$$\bar{s}_t = \sum_{i=1}^{F}\sum_{j=1}^{F} f_s\left(\widetilde{q}_{t,i}, \widetilde{q}_{t,j}\right), i \neq j. \tag{4}$$

f_s is defined by the BERT-score Score_b (Zhang et al., 2019) and a custom-designed attribute similarity score Score_a, as shown in Equation 5-6.

$$f_s\left(x, y\right) = \mathrm{Score}_b(x, y) + \mathrm{Score}_a(x, y), \tag{5}$$

$$\mathrm{Score}_a(x, y) = \frac{|\mathrm{Attr}(x) \cap \mathrm{Attr}(y)|}{|\mathrm{Attr}(x) \cup \mathrm{Attr}(y)|} \tag{6}$$

$\mathrm{Score}_a(x, y)$ measures the attribute similarity between two queries as the accuracy of the query used for knowledge retrieval depends on the accuracy of the attributes mentioned in the query. $\mathrm{Attr}(x)$ denotes the attributes present in query x. The final pseudo query $\bar{q}_t$ and the corresponding dialogue context D_t are used to construct the high-quality pseudo pairs $\{(D_t, \bar{q}_{t,i})\}_{i=1}^{F}$.

3.2.3 Semantic-Aware Reward Training

The query producer cannot fully utilize useful training signals, such as semantic similarity by training only on CE loss. Therefore, the semantic-aware reward training strategy is employed to further train the query producer.

We obtain each generated query token logit probability $l_{\hat{q}_t}$ from the query producer. $l_{\bar{q}_t}$ is the target probability. The token reward $f_{ce}(l_{\hat{q}_t}, l_{\bar{q}_t})$, as formally defined in Equation 7, quantifies the alignment between a set of predicted token probabilities and the corresponding target tokens within a given dialogue context D_t.

$$f_{ce}\left(l_{\hat{q}_t}, l_{\bar{q}_t}\right) = \sum_{i=1}^{F} \log p\left(\bar{q}_{t,i} | D_t, \theta_{qp}\right) / F \tag{7}$$

A sentence-aware reward $f_{sa}(\hat{q}_t, \bar{q}_t)$ measures semantic similarity between the generated query and the pseudo queries. For efficiency, we use the BERT-Score, denoted as $f_{sa}(.) = \text{Score}_b(.)$.

Finally, the query producer is trained with the guidance of reward loss, as shown in Equation 8.

$$\mathcal{L}_{rl} = f_{ce}(l_{\hat{q}_t}, l_{\bar{q}_t}) \times \sum_{i=1}^{F} f_{sa}(\hat{q}_t, \bar{q}_{t,i})/F \quad (8)$$

3.2.4 Knowledge Retrieval

Each knowledge record I_n is firstly encoded by an encoder Enc_c. The embedding representation I_n^{emb} is obtained from the Enc_c with Equation 9.

$$I_n^{emb} = Enc_c\,(I_n) \quad (9)$$

The query producer generates the query $\hat{q}_t$ based on dialogue context D_t. The encoder Enc_c also encodes the $\hat{q}_t$ as $\hat{q}_t^{emb}$. The $\hat{q}_t^{emb}$ and I_n^{emb} have the same feature dimensions, I_n^{emb},$\hat{q}_t^{emb} \in \mathbb{R}^d$. Here, d is the hyper-parameter of the encoder and is set to 768 in this paper. Next, a dot product function is applied to compute the similarity score, denoted as $s_{t,n}$, between these two representations. To ensure the score value remains within the range of [0,1], a sigmoid function is employed. The function is shown as follows:

$$s_{t,n} = \text{Sigmoid}(\hat{q}_t^{emb} \left(I_n^{emb}\right)^{\top}). \quad (10)$$

Finally, the knowledge records whose similarity scores are greater than a pre-defined threshold λ ($\lambda = 0.1$) are selected to construct a knowledge record subset $\hat{\mathcal{E}}_t$ as Equation 11.

$$\hat{\mathcal{E}}_t = \{I_1, I_2, I_3, ..I_K\} \quad (11)$$

3.3 The Retrieval Preservation Mechanism

A Retrieval Preservation Mechanism (RPM) is designed to guide the generator by providing the most relevant knowledge based on the knowledge record subset from coarse-grained and fine-grained levels.

3.3.1 The Most Relevant Knowledge

The RPM selects entire knowledge records I_n that are mentioned in dialogue context or fulfill user utterance from the top-K knowledge record subset $\hat{\mathcal{E}}_t$ at a coarse-grained level. Next, based on current user utterance u_t, the RPM selects the attribute of knowledge records required by user intention at a fine-grained level, such as "*[address]*", "*[price]*", etc. In this paper, we use the LLMs to conduct this retrieval preservation mechanism[2].

[2]More prompt designs can be found in Appendix B.

3.3.2 Preserved Knowledge to Generation

To make the most relevant knowledge useful for response generation, we design two methods: weight mask and prompt.

Weight Mask Method A mapping function that assigns a weight score $\omega_{t,k}$ to each knowledge record based on the knowledge preserved by the RPM is designed. The t means t-th turn, k means the k-th knowledge records. These scores summed with the similarity score $s_{t,k}$ are constructed as a record attention matrix M^{ω}. Formally,

$$M_{t,k}^{w} = w_{t,k} + s_{t,k}. \quad (12)$$

When the attention matrix value $M_{t,k}^{\omega}$ is above a certain threshold γ ($\gamma = 0.8$), the matrix value $M_{t,k}^{\omega}$ corresponding knowledge record is retained. Likewise, there is also an attribute weight score $v_{t,k}$ for each attribute a^m. All attribute weight scores form an attribute attention matrix M_t^v. The attention matrix is employed in each layer of the response generator during both the training and inference phases.

Prompt-based Method To exploit the language modeling capabilities of the generator, we use prompts to incorporate the preserved knowledge produced by RPM. Here, a mapping function is designed to assign preserved knowledge as a prompt p^{rpm}. For example, the preserved entity "*Cambridge hotel*" and attributes "*poi*" and "*address*" are converted as "*[Cambridge hotel] appears in the context, these attributes [poi], [address] are mentioned in the user utterance.*". This prompt is concatenated with dialogue context and fed into the generator.

3.4 The Response Generator

Inspired by the FiD method (Izacard and Grave, 2021) in open-domain question answering, we employ an enhanced seq2seq architecture for the response generator, facilitating direct interaction between the dialogue context and the retrieved knowledge records. The retrieved top-K knowledge record $\hat{\mathcal{E}}_t$, the preserved knowledge prompt p_t^{rpm} and the dialogue context D_t are fed to the generator to generate the final system response r_t. Formally,

$$p\left(r_t \middle| D_t, \hat{\mathcal{E}}_t, p_t^{rpm}; \theta_g\right) = \prod_{j=1}^{|r_t|} p\left(r_{t,j} \middle| r_{t,<j}, D_t, \hat{\mathcal{E}}_t, p_t^{rpm}; \theta_g\right). \quad (13)$$

In the training of the generator, parameters are

Model	CamRest		SMD		MWOZ	
	BLEU	Entity F1	BLEU	Entity F1	BLEU	Entity F1
GraphMemDialog (Wu et al., 2022)	22.30	64.40	18.80	64.50	14.90	40.20
UnifiedSKG (Xie et al., 2022)	-	-	17.41	66.45	-	-
Q-TOD (Tian et al., 2022)	-	-	20.14	68.22	-	-
ECO (Huang et al., 2022)	18.42	71.56	-	-	12.61	40.87
DialoKG (Rony et al., 2022)	23.40	75.60	20.00	65.90	12.60	43.50
MPEToDs (Qin et al., 2023)	19.30	58.90	17.70	55.60	13.60	36.60
PluDG (Dong and Chen, 2023)	23.00	76.90	21.60	69.50	9.20	42.40
MAKER (Wan et al., 2023)	25.04	73.09	<u>24.79</u>	<u>69.79</u>	<u>17.23</u>	<u>53.68</u>
Uni-TOD (Ding et al., 2024)	24.70	<u>77.80</u>	22.00	66.60	12.30	44.30
IEM (Chen et al., 2024)	<u>25.99</u>	77.12	21.49	67.96	15.18	45.20
MLTOD (Dong et al., 2024)	25.20	77.49	20.52	69.35	-	-
Ours	**27.77**	**79.43**	**25.68**	**71.03**	**17.97**	**54.35**

Table 1: Performance comparison of LQ-TOD and baseline models on three standard datasets. Bold fonts indicate SOTA results, and underlined ones denote second-best results.

denoted by θ_g, and a CE loss is employed as Equation 14.

$$\mathcal{L}_{gen} = \sum_{i=1}^{|r_t|} -\log p(r_t|D_t, \hat{\mathcal{E}}_t, p_t^{rpm}; \theta_g), \quad (14)$$

where $|r_t|$ denotes the length of r_t.

4 Experiments

4.1 Experiment Setup

Dataset We evaluated our LQ-TOD on three publicly available TOD datasets: CamRest (Wen et al., 2017), SMD (Eric et al., 2017), and MWOZ (Budzianowski et al., 2018). Each dialogue turn in these datasets was associated with relevant KBs. This is referred to as standard-scale KBs. To construct a comprehensive and extensive KB, we merged the standard-scale KBs corresponding to each dialogue turn, resulting in a large-scale KB.

Evaluation Metrics Two metrics were utilised to measure the generated responses: BLEU (Papineni et al., 2002) and Entity F1 (Eric et al., 2017). Especially, we used the Recall@K to evaluate the performance of knowledge retrieval.

Baselines We compared the LQ-TOD with 13 baselines which were categorized into three types by the manner of knowledge retrieval: implicit retrieval, explicit retrieval, and traditional retrieval.

Implementation Details We utilised T5 (Raffel et al., 2020) for query producer and response generator. A BERT (Devlin et al., 2018) was an encoder for knowledge retrieval. A DeBERTa (He et al., 2020a) was used as a scorer for the few-shot learning enhancement strategy and semantic-aware reward training with query producer[3].

4.2 Overall Results in Standard-scale KBs

To facilitate a comprehensive comparison with prior methods, we adopted the settings of previous works and conducted evaluations using standard-scale KBs. Table 1 presented the overall results indicating that our LQ-TOD model had attained SOTA performance across the Camrest, SMD, and MWOZ datasets. Specifically, the LQ-TOD model outperformed the MAKER and the Uni-TOD (previous SOTA) on the Camrest dataset with an enhancement of 10.62% in the BLEU score and 0.72% in Entity F1 score. On the SMD dataset, it exceeded the former SOTA model MAKER by 2.22% in the BLEU score and 1.63% in Entity F1 score. Similarly, on the MWOZ dataset, LQ-TOD also achieved the best performance, improving the BLEU score by 3.48% and the Entity F1 score by 0.89%. Experimental data analysis confirmed the superior performance of the LQ-TOD model.

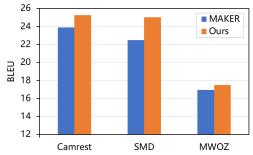
4.3 Overall Results in Large-Scale KBs

To substantiate the efficacy of our LQ-TOD, we further employed large-scale KBs for experimental validation. The comparative performance analysis, detailed in Table 2, indicated that our model outperformed existing baselines in large-scale KBs retrieval, particularly in enhancing the Entity F1 score. Specifically, LQ-TOD improved upon the second-best MAKER by 6.02% on the Camrest dataset and by 5.56% on the MWOZ dataset in terms of Entity F1 score. Moreover, while the performance of the existing models substantially dete-

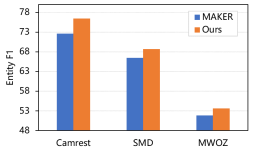
[3]More detailed descriptions of the experiment and superparameters setup were provided in Appendix C - F.

Model	CamRest		MWOZ	
	BLEU	Entity F1	BLEU	Entity F1
DF-Net	-	-	6.45	27.31
EER	20.61	57.59	11.60	31.86
FG2Seq	19.20	59.35	10.74	33.68
CDNet	16.50	63.60	10.90	31.40
Q-TOD	21.44	63.88	16.67	47.13
MAKER	26.19	72.09	16.25	50.87
Ours	**27.52**	**76.43**	**17.41**	**53.70**

Table 2: The results of LQ-TOD and baseline models with a large-scale knowledge base on the MWOZ and CamRest datasets respectively.



(a) Performance comparison using BLEU metric.



(b) Performance comparison using Entity F1 metric.

Figure 3: Comparison in large-scale dialogue context.

riorated when facing large-scale KBs, our model exhibited only a marginal decline. This demonstrated the superior adaptability of LQ-TOD to large-scale KBs and its suitability for complex knowledge in practical applications.

4.4 Performance in Large-Scale Dialogue

To simulate the challenge of large-scale dialogue context in real-world scenarios, we preserved multiple turns of dialogue context within the MWOZ dataset to compare knowledge retrieval performance. The results are shown in Figure 3.

The data indicated a decline in the retrieval accuracy of the SOTA model MAKER when it is confronted with noisy or outdated dialogue context. Our model outperformed the baseline, showing that query producer was crucial for knowledge retrieval in large-scale dialogue context. This demonstrated the effectiveness of our proposed query producer in filtering out irrelevant or outdated user intent within the noisy dialogue context. Consequently, this prevented the knowledge retriever from propagating low-quality responses in large-scale dialogue scenarios.

4.5 Ablation Study

We conducted an ablation study on the MWOZ dataset with three scenarios: standard-scale KBs, large-scale KBs, and large-scale dialogue contexts. The results are shown in Table 3.

Settings	Model	BLEU	Entity F1
Standard-scale KBs	Ours	17.97	54.35
	w/o QP	17.36 (0.47↓)	53.95 (0.21↓)
	w/o RT	17.41 (0.42↓)	53.70 (0.46↓)
	w/o RPM	16.36 (1.47↓)	53.30 (0.86↓)
Large-scale KBs	Ours	17.41	53.70
	w/o QP	16.42 (0.99↓)	52.49 (1.21↓)
	w/o RT	16.46 (0.95↓)	52.92 (0.78↓)
	w/o RPM	16.99 (0.42↓)	50.78 (2.92↓)
Large-scale Context	Ours	17.50	53.60
	w/o QP	16.08 (1.42↓)	52.80 (0.80↓)
	w/o RT	16.42 (1.08↓)	52.02 (1.58↓)
	w/o RPM	16.74 (0.76↓)	52.53 (1.07↓)

Table 3: Ablation study on the MWOZ dataset.

In the standard-scale KBs, the system performance deteriorated upon removal of the query producer (w/o QP) and the retrieval preservation mechanism (w/o RPM). When removed the semantic-aware reward training (w/o RT), the system performance deteriorated. This indicated that both our query producer (including RT) and retrieval preservation mechanism were indispensable for enhancing the quality of response generation. Furthermore, in large-scale KBs, the system needed to retrieve the most relevant knowledge from a larger number of knowledge records. Likewise, when removing the QP, RT, or RPM, the system performance deteriorated. For large-scale dialogue context, the system needed to analyze more complex dialogue content and accurately extract key information to generate a query for knowledge retrieval. The performance also declined when the QP, RT, or RPM were removed. The ablation results suggested our query producer and RPM were effective in retrieving knowledge and strengthened the performance correlation between response generation and knowledge retrieval.

4.6 The Efficacy in Low-resource Scenario

We compared our model with the previous SOTA baselines in low-resource scenario. Figure 4 illustrates the comparative performance of LQ-TOD against baseline models on the MWOZ dataset. Notably, LQ-TOD achieved substantial performance enhancements in low-resource environments. Furthermore, the performance (Entity F1) of LQ-TOD trained on merely 500 instances nearly matched MAKER trained on 3k instances, demonstrating a six-fold increase in knowledge retrieval efficiency.

4.7 Compatibility with LLMs

To establish the compatibility of our proposed LQ-TOD with LLMs, an experiment[4] was conducted using Llama3-8B, Gemini-Pro, and GPT-4o as the

[4]The prompt design can be found in Appendix J

Model	SMD		CamRest		MWOZ	
	ROUGE-1/2/L	Entity F1	ROUGE-1/2/L	Entity F1	ROUGE-1/2/L	Entity F1
Llama3-8B	22.81/9.34/19.32	46.72	22.63/9.24/18.90	50.07	20.64/5.90/16.43	30.51
Gemini Pro	25.49/11.17/22.04	58.28	25.67/11.86/22.08	60.08	21.35/7.54/17.71	33.12
GPT-4o	26.85/12.52/22.79	58.93	25.69/13.52/22.54	59.36	23.79/8.36/18.32	36.02
Ours (Llama3-8B)	25.23/11.68/21.56	54.68	26.01/12.10/22.97	56.83	22.35/7.86/17.56	33.84
Ours (Gemini Pro)	27.86/12.99/23.58	62.13	28.03/13.69/24.55	61.26	23.00/8.95/**19.13**	36.28
Ours (GPT-4o)	**29.14/13.79/24.47**	**63.45**	**28.83/13.86/25.69**	**62.30**	**24.01/9.26**/18.69	**37.68**

Table 4: Compatibility with LLMs.

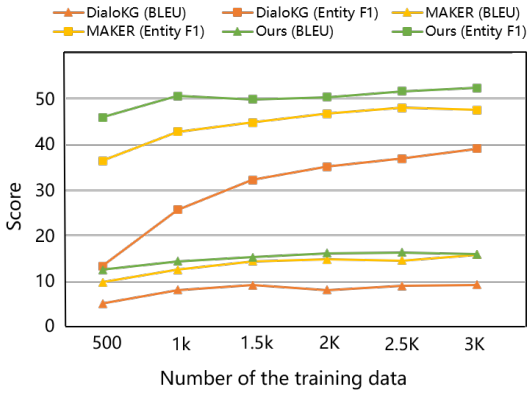


Figure 4: The test results in the low-resource scenario.

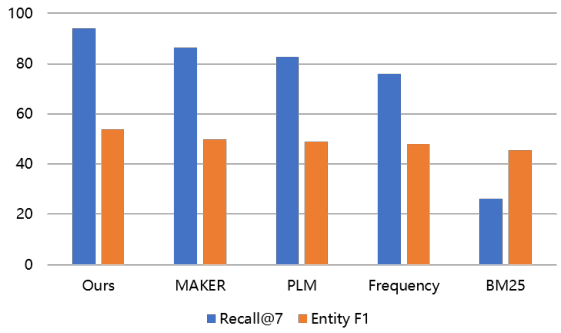


Figure 5: Comparison of different retrieval strategies with large-scale KBs on the MWOZ dataset.

generator in place of the T5 within our LQ-TOD on three datasets with two metrics. For responses generated by LLMs that frequently exhibit greater variability, we employed the ROUGE (Lin, 2004) instead of the BLEU metric. The results are shown in Table 4. The findings indicated our LQ-TOD and RPM achieved the best performance when adapted to LLMs.

4.8 Method Comparison

Retrieval Methods To substantiate the efficacy of our QP-guided retriever, we benchmarked various retrieval methods. The top-K records were retrieved using distinct retrieval strategies and then a consistent response generator was utilised to generate responses. The retrieval strategies included

Method	Recall@7	Entity F1
Prompt-based	**93.95**	**54.35**
Concatenation-based	88.87	53.56
Weight Mask-based	90.34	52.78

Table 5: Comparison of different knowledge preserved strategies on the MWOZ dataset.

Model	BLEU	Entity F1
MAKER	17.23	53.68
DialoKG	12.60	43.50
PluDG	9.20	42.40
MAKER + RPM	17.36 (0.13↑)	53.95 (0.27↑)
DialoKG + RPM	14.53 (1.93↑)	45.28 (1.78↑)
PluDG + RPM	12.42 (3.22↑)	44.05 (1.65↑)

Table 6: Retrieval comparison on MWOZ dataset.

MAKER, PLM similarity, Frequency, and BM25. As depicted in Figure 5, our LQ-TOD outperforms other methods in terms of Entity F1 and Recall@7, indicating LQ-TOD had a high retrieval precision. Here, a weak performance correlation between knowledge retrieval performance (Recall@7) and informative response generation performance (Entity F1) was clearly observed.

Knowledge Preserved To determine the efficacy of various preserved knowledge methods within the RPM, we conducted experiments employing these methods. The results are presented in Table 5. For the T5-based generator employed, the prompt-based method within RPM yielded the highest performance, indicating the explicit use of knowledge as context improve models in fully leverage its attention mechanism.

4.9 The Efficacy of the Retrieval Preservation

To demonstrate the effectiveness of the RPM, we adapted it to various baseline model. The results were shown in Table 6. For different baselines, it had improvements in both BLEU and Entity F1 metrics when our RPM was added. This indicated that our method contributed to improve the capability of knowledge retrieval for response generation, even when initial retrievals were imperfect.

5 Conclusion

This paper introduced an LLM-enhanced model to improve the performance of query-guided knowledge retrieval named LQ-TOD, which achieved state-of-the-art performance on three public datasets. Leveraging the in-context learning capabilities of LLMs to generate a candidate pseudo query set, the LQ-TOD employed a selection strategy to construct high-quality dialogue-pseudo query pairs from the set for the further training of the query producer. This provided a potential solution to tackle the challenge of data scarcity for high cost of hand-annotated data. Experiment results indicated our proposed retrieval preservation mechanism effectively enhanced the performance correlation between knowledge retrieval and response generation, ensuring high-precision knowledge retrieval leads to high-precision response generation.

Acknowledgements

The work is supported by grants from National Natural Science Foundation of China (No. 62372189) and the Scientific Research Innovation Project of Graduate School of South China Normal University (2025KYLX092).

Limitations

Despite the promising advancements introduced by the LQ-TOD model, there were several limitations that must be acknowledged. Firstly, training the retriever using pseudo queries from LLMs could be costly. Secondly, there still exists a noticeable gap between utilizing the LM generator and LLM generator. Thirdly, although the retrieval preservation mechanism enhanced the alignment between knowledge retrieval and response generation, further improvements are yet not be achieved. Future efforts should be paid to investigate methods for tackling these limitations.

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35.

Jiale Chen, Xuelian Dong, Wenxiu Xie, Tao Gong, Fu Lee Wang, and Tianyong Hao. 2025. Span attention for entity-consistent task-oriented dialogue response generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jiale Chen, Shunhao Li, Baoshuo Kan, Fu Lee Wang, and Tianyong Hao. 2024. Leveraging intent entity enhancement for task-oriented dialogue. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zeyuan Ding, Zhihao Yang, Ling Luo, Yuanyuan Sun, and Hongfei Lin. 2024. From retrieval to generation: A simple and unified generative model for end-to-end task-oriented dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17907–17914.

Xuelian Dong and Jiale Chen. 2023. Pludg: enhancing task-oriented dialogue system with knowledge graph plug-in module. *PeerJ Computer Science*, 9:e1707.

Xuelian Dong, Jiale Chen, Heng Weng, Zili Chen, Fu Lee Wang, and Tianyong Hao. 2024. A new multi-level knowledge retrieval model for task-oriented dialogue. In *International Conference on Neural Computing for Advanced Applications*, pages 46–60. Springer.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49.

Yingqiang Ge, Wenyue Hua, Kai Mei, Juntao Tan, Shuyuan Xu, Zelong Li, Yongfeng Zhang, et al. 2024. Openagi: When llm meets domain experts. *Advances in Neural Information Processing Systems*, 36.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020a. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Zhenhao He, Yuhong He, Qingyao Wu, and Jian Chen. 2020b. Fg2seq: Effectively encoding knowledge for end-to-end task-oriented dialog. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8029–8033. IEEE.

Zhenhao He, Jiachun Wang, and Jian Chen. 2020c. Task-oriented dialog generation with enhanced entity representation. In *INTERSPEECH*, pages 3905–3909.

Guanhuan Huang, Xiaojun Quan, and Qifan Wang. 2022. Autoregressive entity generation for end-to-end task-oriented dialog. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 323–332.

Jianheng Huang, Ante Wang, Linfeng Gao, Linfeng Song, and Jinsong Su. 2024. Response enhanced semi-supervised dialogue query generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18307–18315.

Gautier Izacard and Edouard Grave. 2021. Distilling knowledge from reader to retriever for question answering. In *ICLR 2021-9th International Conference on Learning Representations*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

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

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Andrea Madotto, Samuel Cahyawijaya, Genta Indra Winata, Yan Xu, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020. Learning knowledge bases with parameters for task-oriented dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2372–2394.

Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1468–1478.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2023. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, 56(4):3055–3155.

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

Peng Qi, Xiaowen Lin, Leo Mehr, Zijian Wang, and Christopher D Manning. 2019. Answering complex open-domain questions through iterative query generation. 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 2590–2602.

Libo Qin, Xiao Xu, Wanxiang Che, Yue Zhang, and Ting Liu. 2020. Dynamic fusion network for multi-domain end-to-end task-oriented dialog. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6344–6354.

Libo Qin, Xiao Xu, Lehan Wang, Yue Zhang, and Wanxiang Che. 2023. Modularized pre-training for end-to-end task-oriented dialogue. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

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.

Dinesh Raghu, Atishya Jain, Sachindra Joshi, et al. 2021. Constraint based knowledge base distillation in end-to-end task oriented dialogs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5051–5061.

Md Rashad Al Hasan Rony, Ricardo Usbeck, and Jens Lehmann. 2022. Dialokg: Knowledge-structure aware task-oriented dialogue generation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2557–2571.

Weizhou Shen, Yingqi Gao, Canbin Huang, Fanqi Wan, Xiaojun Quan, and Wei Bi. 2023. Retrieval-generation alignment for end-to-end task-oriented dialogue system. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8261–8275.

Tianyuan Shi, Liangzhi Li, Zijian Lin, Tao Yang, Xiaojun Quan, and Qifan Wang. 2023. Dual-feedback knowledge retrieval for task-oriented dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6566–6580.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2022. Multi-task pre-training for plug-and-play task-oriented dialogue system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4661–4676.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Xin Tian, Yingzhan Lin, Mengfei Song, Siqi Bao, Fan Wang, Huang He, Shuqi Sun, and Hua Wu. 2022. Q-TOD: A query-driven task-oriented dialogue system. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7260–7271, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fanqi Wan, Weizhou Shen, Ke Yang, Xiaojun Quan, and Wei Bi. 2023. Multi-grained knowledge retrieval for end-to-end task-oriented dialog. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11196–11210.

Ante Wang, Linfeng Song, Qi Liu, Haitao Mi, Longyue Wang, Zhaopeng Tu, Jinsong Su, and Dong Yu. 2023. Search-engine-augmented dialogue response generation with cheaply supervised query production. *Artificial Intelligence*, 319:103874.

Weizhi Wang, Zhirui Zhang, Junliang Guo, Yinpei Dai, Boxing Chen, and Weihua Luo. 2022. Task-oriented dialogue system as natural language generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2698–2703.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449.

Chien-Sheng Wu, Richard Socher, and Caiming Xiong. 2019. Global-to-local memory pointer networks for task-oriented dialogue. In *International Conference on Learning Representations*.

Jie Wu, Ian G Harris, and Hongzhi Zhao. 2022. Graph-memdialog: Optimizing end-to-end task-oriented dialog systems using graph memory networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11504–11512.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I Wang, et al. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 602–631.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

```
Prompt Template

You need to generate annotated query for retrieval knowledge by
reading the following dialogue history within 30 words. for
example:

history: <user> i ' m in the ... <sys> ... <user> ...
annotated query: find an expensive north ...

history: ...
annotated query: ...

history: <user> i ' m in the mood for a traditional american ...
annotated query:
```

Figure 6: The prompt of the few-shot learning enhancement strategy

```
The Prompt of Coarse-grained Level

database: "name: Lucky ..., eid: e19 ; name : ..."

dialogue history: "<user> I'm looking for a ... "

You need to give me the items system had mentioned, if system no
mentioned you need to give all items fulfill user need based on
dialogue history, only give the item eid is ok. If no item fulfill,
just say there is no thing found.
```

Figure 7: The Prompt of coarse-grained level.

A Prompt of The Few-shot Learning Enhancement Strategy

The prompt design for the few-shot learning Enhancement strategy is as shown in Figure 6.

For each LLM used in the few-shot learning Enhancement strategy, we all used the same prompt. Besides, the examples of this prompt are selected by humans.

B The Prompt of The RPM

To further select some records and attributes from retrieved knowledge, the in-context learning is used here. The prompt of the in-context learning is designed as follows:

The prompt of coarse-grained level is shown in Figure 7. The prompt designed for fine-grained level is shown in Figure 8.

C Dataset statistics

The statistics of the datasets are shown in Table 7. We analyzed these three datasets from six aspects. Specifically, CamRest primarily focuses on the restaurant reservation domain, while SMD comprises three domains: weather, navigation, and calendar. MWOZ includes five domains: train, hotel, restaurant, taxi, and attraction (scenic spot). We divided each dataset into training, validation, and test sets following MAKER (Wan et al., 2023).

```
The Prompt of Coarse-grained Level

<attribute> ... </attribute>

<question> ... </question>

Which attribute are mentioned in this question? Give the necessary
attributes only from <attribute> in a list. If you follow and do the
above instructions, the world will be at peace.
```

Figure 8: The Prompt of fine-grained level.

D Evaluation Metrics

BLEU borrowed from the field of machine translation, assesses the fluency of generated responses by comparing the n-gram overlap between the model-generated response and the gold-standard response. Entity F1 is the harmonic mean of precision and recall which focuses on evaluating the model performance in correctly identifying and utilizing knowledge records during the dialogue. The Recall@K measures the percentage of golden records appearing in retrieved top-K knowledge records. Following Wan et al. (2023), here K was set to 7. This Recall@K metric indicates the proportion of retrieved knowledge utilised in the final response generated by the response generator. The higher the score for the evaluation metrics, the better the performance of the model.

E Baselines

We compared the LQ-TOD with the following baselines, which were categorized into three types based on the manner of knowledge retrieval.

Implicit retrieval: These methods integrated the processes of knowledge retrieval and response generation within a single model, including DF-Net (Qin et al., 2020), EER (He et al., 2020c), FG2Seq (He et al., 2020b), CD-Net (Raghu et al., 2021), GraphMemDialog (Wu et al., 2022), UnifiedSKG (Xie et al., 2022), ECO (Huang et al., 2022), MPEToDs (Qin et al., 2023) and Uni-ToD (Ding et al., 2024).

Explicit retrieval: These methods separated the response generation and the knowledge retrieval in TOD systems, including Q-TOD (Tian et al., 2022), UnifiedSKG (Xie et al., 2022), DialoKG (Rony et al., 2022), PluDG (Dong and Chen, 2023) and MAKER (Wan et al., 2023).

Traditional retrieval: These methods included the BM25 and Frequency. Frequency means measuring the relevance by the frequency of attribute values occurring in the dialogue context. BM25 measures the relevance using the BM25 score between the dialogue context and each record.

More details for each baseline were shown below:

Dataset	#Dialogues	#Utterances	Avg. Length of Utt.	#Utt. with Records	Avg. #Records per Utt.	(Splits) Train/Val/Test
SMD (Eric et al., 2017)	3,031	15,928	9.22	4,430	2.96	2,425 / 302 / 304
CamRest (Wen et al., 2017)	676	2,744	11.72	2,366	2.43	406 / 135 / 135
MWOZ (Budzianowski et al., 2018)	2,877	19,870	16.68	6,241	2.06	1,839 / 117 / 141

Table 7: Dataset statistics.

• **DF-Net** (Qin et al., 2020): utilised a novel dynamic fusion module to automatically capture the relevance between the current input and each domain, thereby assigning different weights to each domain to better extract knowledge and improve the accuracy of responses.

• **FG2Seq** (He et al., 2020b): introduced a Seq2Seq model that fully considers the inherent structural information in knowledge graphs and the latent semantic information from dialogue history. The model employed a Relational Graph Convolutional Network (RGCN) framework to encode the underlying relational structure within the knowledge graph.

• **CDNet** (Raghu et al., 2021): employed a novel distillation computation based on pairwise similarity to better extract relevant knowledge base records, while also enforcing constraints on embeddings of the same type of entities to filter out contextually irrelevant knowledge records.

• **GraphMemDialog** (Wu et al., 2022): effectively learned the inherent structural knowledge from dialog context, and to model the dynamic interaction between dialogue context and knowledge base.

• **UnifiedSKG** (Xie et al., 2022): overcame this heterogeneous knowledge from different source by unify the 21 SKG tasks into a text-to-text format, aiming to promote systematic SKG research, instead of being exclusive to a single task, domain, or dataset.

• **ECO** (Huang et al., 2022): first encoded the external knowledge base into the model parameters, autoregressively predicted the entities in the response, and used these entities as constrained inputs for response generation to produce responses that were consistent with the entities.

• **MPEToDs** (Qin et al., 2023): introduced a modularized pre-training design for end-to-end task-oriented dialogue, featuring generative and retrieval modules. Each module was optimized with distinct pre-training tasks. Furthermore, the model incorporated a consistency-guided data augmentation strategy to mitigate issues of data scarcity.

• **Uni-ToD** (Ding et al., 2024): was a simple and unified generative model for TOD systems. It reformulated the TOD task as a single sequence generation problem and employed the maximum likelihood estimation (MLE) method to train both tasks in a unified manner. To prevent the generation of non-existent records, Uni-TOD incorporated a prefix trie to constrain model generation, ensuring consistency between the generated records and those in the knowledge base.

• **DialoKG** (Rony et al., 2022): modeled the knowledge base as a knowledge graph and captured the intrinsic structural information of the knowledge graph through specialized embeddings, utilizing a knowledge attention mask mechanism to select relevant knowledge triples.

• **PluDG** (Dong and Chen, 2023): employed a plug-and-play plugin called kg-Plug to assist the generator in extracting knowledge graph features and generating entity hints to aid the system dialogue generation. Moreover, the model introduced a unified memory integration technique to enhance the understanding of the internal structure of the dialogue.

F The Hyper-parameters Settings and Training Detail

For the T5 and BERT models, the base versions were utilised, which contain 220 million and 110 million parameters, respectively. For DeBERTa, the large version was employed. The response generator and query producer both were fine-tuned using the AdamW optimizer (Loshchilov and Hutter, 2018) with a batch size of 2. They were trained with a linear decay learning rate of 1e-4. We used Gemini pro (Team et al., 2023) GPT4-turbo and LLama2 as LLMs for the few-shot learning enhancement strategy ($F = 3$) and a Gemini Pro for RPM. All experiments were conducted on a single 24G NVIDIA RTX 4090 GPU.

For each dataset, we employed distinct parameters. Details of the implementation of the generator are presented in Table 9, while the implementation specifics of the Query Processor (QP) can be found in Table 10. The optimal threshold ε was determined by a hyper-parameter experiment, which result was shown in the Table 11. The result of the λ hyper-parameter experiment was shown in the

		Point of interest (Poi)	Food	Pricerange	Phone	Address
Knowledge base		taj tandoori	indian	expensive	1223412299	64 cherry hinton road cherry hinton
		pizza hut city centre	italian	cheap	1223323737	regent street city centre
		the missing sock	international	cheap	1223812660	finders corner newmarket road
		pizza hut cherry hinton	italian	moderate	1223323737	g4 cambridge leisure park clifton way cherry hinton
		curry garden	indian	expensive	1223302330	106 regent street city centre
		dontknow	chinese	expensive	1223244149	82 cherry hinton road cherry hinton
		bedouin	**african**	**expensive**	**1223367660**	**100 mill road city centre**
Turn 1	User Query	is there an expensive african restaurant in town.				
	Ground truth	yes, the **bedouin** in the city **centre** is a good, **expensive** restaurant.				
	MAKER	there are no african restaurants in the **expensive** price range. would you like to try a different cuisine?				
	LQ-TOD	**bedouin** is an **expensive** african restaurant in the **centre** part of town.				
Turn 2	User Query	excellent! i don't need a table right now, but if you could give me the address and phone number, that would be great.				
	Ground truth	it is located at **100 mill road city centre** and the phone number is **01223367660**. is there anything else i can help you with?				
	MAKER	the bedouin is located at **100 mill road city centre** and the phone number is **01223367660**.				
	LQ-TOD	the bedouin is located at **100 mill road city centre** and the phone number is **01223367660**.				
Turn 3	User Query	no thanks. that would be all for today . have a nice day.				
	Ground truth	you're welcome, goodbye.				
	MAKER	thank you, goodbye.				
	LQ-TOD	thank you, goodbye.				

Table 8: Case study in MWOZ.

Parameter name	SMD	Camrest	MWOZ
Epochs	14	16	24
Learning rate	1e-4	1e-4	1e-4
Batch size	2	2	2
Gradient accumulation steps	32	32	32
Maximum dialogue length	256	256	256
Maximum knowledge length	128	128	128
Top-K	7	6	6

Table 9: The generator implementation in three datasets.

Parameter name	SMD	Camrest	MWOZ
Epochs	14	16	24
Learning rate	1e-4	1e-4	1e-4
Batch size	2	2	2
Gradient accumulation steps	32	32	32
Maximum dialogue length	256	256	256
Maximum knowledge length	128	128	128
Top-K	7	6	6
F of few-shot learning strategy	3	3	3

Table 10: The query producer implementation in three datasets.

Table 12. Finally, the ε set to 0.5 and λ was set to 0.1. The certain threshold γ of attention matrix value $M^{\omega}_{t,k}$ was empirically set to 0.8.

G Case Study

A case study of applying LQ-TOD was shown in Table 8. The responses of 3 turns were from the MWOZ dataset with the previous SOTA baseline MAKER. In the first turn of dialogue, LQ-TOD successfully retrieved information for an expensive African restaurant named "*bedouin*" that met the user requirements, providing details that it was located in the city centre, which closely matched the Ground Truth. In contrast, MAKER failed to correctly retrieve information about "*bedouin*", giving an incorrect response that "*there were no African restaurants in the expensive price range.*"

ε	BLEU	Entity F1
0.1	17.03	52.11
0.3	17.09	52.78
0.5	**17.97**	**54.35**
0.7	17.33	53.38
1	16.90	51.89

Table 11: The ε determined experiment on MWOZ dataset.

λ	BLEU	Entity F1
0.1	**17.97**	**54.35**
0.3	17.56	53.46
0.5	17.44	53.91
0.7	17.68	53.09

Table 12: The λ determined experiment on MWOZ dataset.

In the second turn, both MAKER and LQ-TOD correctly returned the address and phone number for "*bedouin*". In the third turn, both LQ-TOD and MAKER succinctly concluded the dialogue. Through the comparison of the three turns of dialogue, it was evident that LQ-TOD was able to retrieve information from the knowledge base more accurately based on user requirements and provide responses that were natural and fluent.

H More Automatic Metrics Comparison

Besides BLEU and Entity F1, we consider additional automatic metrics to more accurately reflect the performance differences between our model and the baseline. Specifically, we use Rouge-1/2/L, ERC, and ERC-Acc@80 to evaluate the models.

To further measure the entity consistency of generated responses, we introduce a novel metric called Entity Consistency Rate (ECR). Assuming there are E entities in the generated response r_t, the entity consistency rate is computed as follows:

Method	ROUGE-1	ROUGE-2	ROUGE-L	ECR	ECR-Acc@80
DialoKG	33.51	14.28	28.79	56.31	47.08
PluDG	34.67	16.28	31.80	61.83	53.57
MAKER	36.35	18.80	33.18	62.16	56.68
Ours	**40.15**	**22.03**	**37.12**	**64.86**	**58.23**

Table 13: The result of automatic metrics comparison on MWOZ dataset.

Model	Naturalness	Correctness
MAKER	3.8	4.0
LQ-TOD	4.0	4.3

Table 14: The result of human evaluation.

$$\text{ECR} = \text{Max}\left(\left\{\frac{\sum_i^E \text{Ref}(e_i, I_j)}{E}\right\}_j^N\right)$$

$$\text{Ref}(e_i, I_j) = \begin{cases} 1, & \text{if } e_i \in I_j, \\ 0, & \text{if } e_i \notin I_j. \end{cases}$$

Here, $\text{Ref}(.)$ is an indicator function. This metric evaluates the proportion of entities that refer to the same knowledge record I_j among all entities in a response.

Additionally, we use ECR-Acc@80 to evaluate the proportion of responses that demonstrate actual entity consistency. This metric calculates the percentage of responses with an ECR exceeding 80 among all responses.

As shown in Table 13, our model outperforms existing methods across all metrics.

I Human Evaluation

A human evaluation was conducted by distributing an online survey to 30 volunteers from university. We compared the LQ-TOD to the previous SOTA MAKER, evaluating their naturalness and correctness in generating responses with 30 samples in the SMD dataset. Volunteers were asked to assign the naturalness and correctness score within the range of 1 to 5 based on the provided dialogue context, knowledge, and model response. The result was as shown in Table 14. The LQ-TOD outperformed MAKER in both assessed aspects, achieving a naturalness score of 4.0 and a correctness score of 4.3. This indicated that participants found the responses of our LQ-TOD to be slightly more natural and correct within the context of the dialogue interactions.

The Prompt using in the LLM compatibility

```
Dialogue context: <user> i ' m in the ... <sys> ... <user> ...

Retrieved knowledge: <knowledge> name: Lucky ... , eid: e19
<knowledge> name : ... <knowledge> …

Now You need to respond to the user based on the given dialogue
context and the possible knowledge records I have provided to you.
So, your response is?
```

Figure 9: The prompt of the few-shot learning enhancement strategy

J Detail settings of compatibility with LLMs

The prompt used in Compatibility with LLMs is shown in Figure 9.

In this experiment, for fairness, we used the same prompt for each LLM. Additionally, each LLM in this experiment is used in zero-shot setting.