SMARTS: A Small-to-large Model Coordinated Retrieval and Response Framework for Task-Oriented Dialogue

Jiale Chen[1], Xuelian Dong[1], Shengcheng Ye[2], Wenxiu Xie[3], Kun Zeng[4], Tianyong Hao[1]*,
[1]School of Computer Science, South China Normal University
[2]School of Computer Science and Technology, East China Normal University
[3]School of Artificial Intelligence, Guangdong Polytechnic Normal University
[4]School of Computer Science, Sun Yat-Sen University
{jlchen, xldong, haoty}@m.scnu.edu.cn, 52285901002@stu.ecnu.edu.cn, wxxie@gpnu.edu.cn, zengkun2@mail.sysu.edu.cn

Abstract

Task-Oriented Dialogue (TOD) systems are commonly used to assist users in achieving specific goals through human-computer interactions. Existing methods typically employ a single model during response generation to simultaneously learn response style and perform fine-grained knowledge retrieval. However, these task-coupling methods often result in suboptimal responses generation and the inclusion of irrelevant knowledge in responses, making the optimization of TOD systems more difficult. To address this challenge, we propose a novel small-to-large model collaboration framework for task-oriented dialogue, named SMARTS. This framework utilizes a small language model to generate style-only responses (without knowledge), while a large language model retrieves top-1 relevant knowledge from the knowledge base independently. The style-only responses are finally filled with top-1 relevant knowledge and form the completed responses to users. Finally, a re-thinker is designed to check the language fluency and knowledge accuracy of responses. Experiments on two public datasets demonstrate that SMARTS outperforms existing methods by an average of 5.79% in Entity F1.

1 Introduction

Task-Oriented Dialogue (TOD) systems are widely used to assist users in fulfilling their information needs through human-computer interaction, such as hotel booking, weather inquiry, and scheduling tasks (Chen et al., 2017). All these tasks require TOD systems to generate highly accurate knowledge-based responses or perform specific actions. Therefore, TOD tasks often necessitate multi-step knowledge retrieval from an external knowledge base to produce accurate user responses (Qin et al., 2023a). Some existing methods directly utilize a single model to generate responses based on both the dialogue context and the entire knowledge base (Rony et al., 2022; Dong and Chen, 2023). In these methods, the Language Model (LM)-based generator produces responses token by token while simultaneously performing knowledge retrieval. However, there are two main challenges: (1) When the knowledge base is large, the model input becomes excessively long, leading to increased response generation time. (2) Since an entire knowledge base is used as input, the model struggles to produce precise responses among a vast number of knowledge records.

To address these challenges, some works propose alleviation strategies (Dong et al., 2026; Chen et al., 2025b; Wan et al., 2023). These methods employ a knowledge retriever to first select the top-K relevant knowledge records from the knowledge base, and feed only these records, along with the dialogue context, into the generator for response generation. However, they still require the LM-based generator to retrieve fine-grained knowledge from the top-K records during response generation. In practice, these methods often generate suboptimal responses that include irrelevant knowledge. Another key limitation of these approaches lies in their retrievers: they typically encode each entire knowledge record into a single vector and perform retrieval based on vector similarity. This coarse-grained matching strategy makes it difficult to capture subtle differences between knowledge records. As a result, the retrieved top-K knowledge may still contain irrelevant or noisy entries. The differences and similarities between these methods are illustrated in Figure 1.

In this paper, we introduce SMARTS (Small-to-large Model Collaboration Framework for Task-Oriented Dialogue), a novel framework designed to alleviate the task-coupling limitations of existing TOD systems. SMARTS fully decouples knowledge retrieval from response generation, enabling each component to focus on a single task. A fine-

*Corresponding author.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 2603–2612
November 6-10, 2026 ©2026 Association for Computational Linguistics

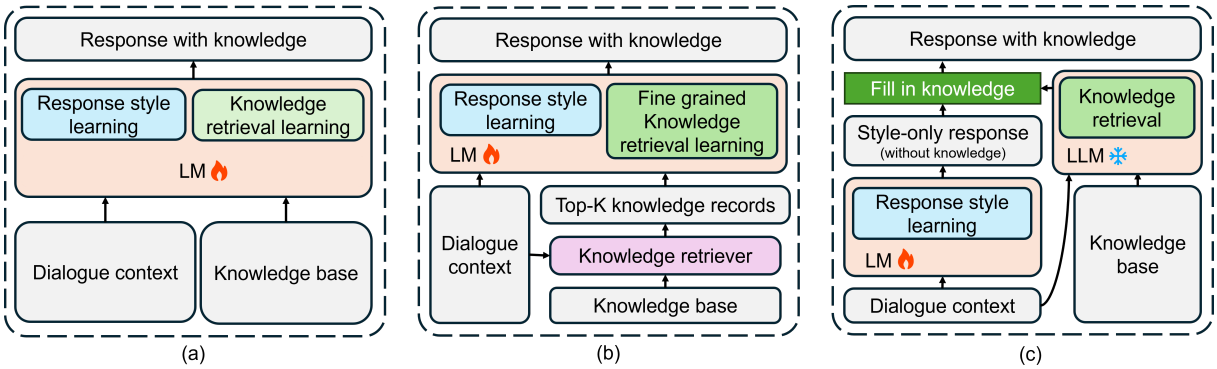


Figure 1: Methods for task-oriented dialogue. (a) This kind of method employs the dialogue context and entire knowledge base as input to LM for response generation with knowledge. (b) This kind of method utilizes a knowledge retriever to conduct coarse grained knowledge retrieval. (c) Our method, which totally decouples response generation and knowledge retrieval, enabling each component to focus on a single task.

tuned Small Language Model (SLM) is trained to generate style-only responses containing placeholder slots for attributes, while a training-free Large Language Model (LLM) leverages its strong reasoning and semantic understanding capabilities to retrieve the most relevant knowledge record (top-1) from an attribute-filtered knowledge base. This knowledge base is dynamically constructed through a fine-to-coarse retrieval strategy that filters irrelevant attributes based on the slot information in the style-only response. Then, the attributes of the top-1 knowledge record are directly filled into the slots of the style-only response. This separation allows for targeted optimization and better generalization. To ensure a fluent and accurate final response, we introduce a re-thinker module that evaluates both the fluency and correctness of the generated output. Experiments on two public datasets demonstrate that SMARTS consistently outperforms existing methods, achieving an average improvement of 5.79% in Entity F1 score. In summary, this paper provides three contributions:

- We propose a new small-to-large model Collaboration framework for task-oriented dialogue named SMARTS, synergizing the efficiency of small language model for response style learning with the precise knowledge retrieval of a frozen large language model, enabling a balanced and adaptable method to task-oriented dialogues.
- We introduce a novel fine-to-coarse knowledge retrieval strategy. It removes the irrelevant knowledge attributes of knowledge base before retrieving the top-1 knowledge record. This strategy enhances the precision in retrieving relevant knowledge while alleviating high computational costs.
- Experiments on all two publicly available datasets demonstrate our proposed method outperform state-of-the-art baselines. A series of in-depth experiments analysis validates the effectiveness of both the small-to-large model collaboration framework and the fine-to-coarse knowledge retrieval strategy.

2 Related work

Existing end-to-end task-oriented dialogue (TOD) systems can generally be categorized into two paradigms: the unified paradigm and the pseudo-pipeline paradigm.

The unified paradigm employs a single model to directly generate responses based on both the dialogue context and the knowledge base. For instance, Madotto et al. (2018) propose Mem2Seq, an end-to-end neural network that uses a memory network to store dialogue context and knowledge records, followed by a pointer network to generate responses. Qin et al. (2020) extend Mem2Seq with a shared-private framework to separately learn domain-shared and domain-specific knowledge. With the rise of pre-trained language models (PLMs), many works adopt PLMs to build unified systems. Budzianowski and Vulić (2019) propose fine-tuning PLMs for dialogue tasks, while Yang et al. (2021) directly use GPT-2 to model the entire dialogue session as a sequence of user utterances, belief states, database results, system acts, and responses. Wang et al. (2022) enhance GPT-2 with lightweight adapter and CopyNet modules (GPT-Adapter-CopyNet) for better domain adaptation and entity generation. Rony et al. (2022) propose DialoKG, which dynamically embeds knowl-

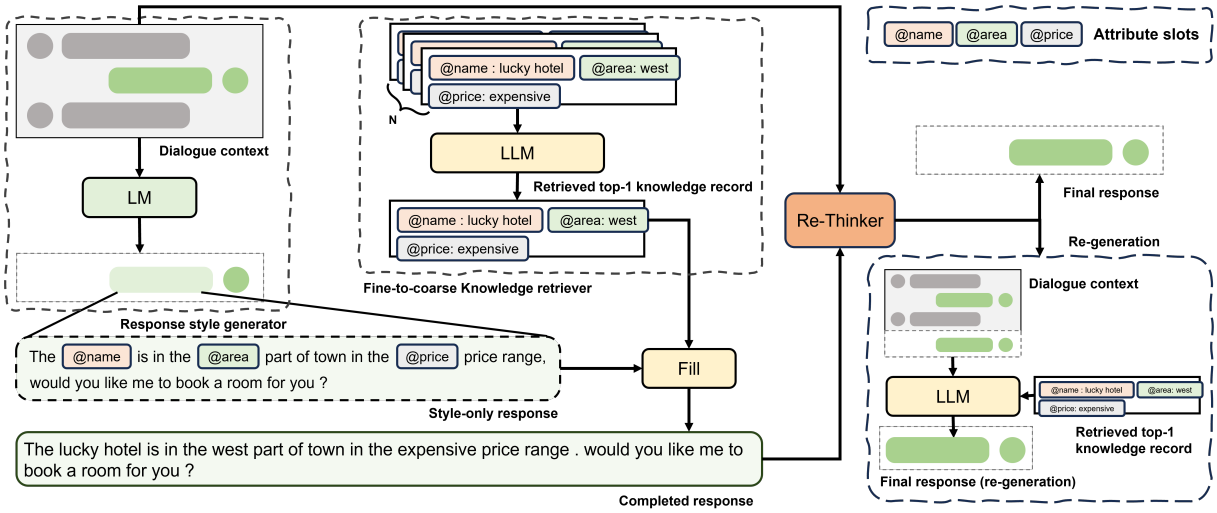


Figure 2: The overall architecture of SMARTS.

edge into context for GPT-2. To further improve generation accuracy, Ding et al. (2024) introduce a prefix trie to constrain GPT-2's outputs, ensuring consistency with the knowledge base.

However, jointly optimizing knowledge retrieval and response generation in a single model often leads to suboptimal performance and poor interpretability. To address this, the pseudo-pipeline paradigm separates the TOD task into two loosely coupled components: a retriever and a generator. For example, Cai et al. (2019) utilize a retriever to retrieve past similar dialogues and extract a retrieval-based skeleton, which is then used to generate a final response. Tian et al. (2022) propose Q-TOD, which extracts query representations from dialogue context for knowledge retrieval, followed by response generation. Wan et al. (2023) propose MAKER, a multi-granular retriever that includes both entity and attribute selectors to improve knowledge relevance. Shen et al. (2023) use maximal marginal likelihood to train the retriever using feedback from response generation. Similarly, Shi et al. (2023) propose a dual feedback mechanism to jointly optimize the retriever and generator. Dong et al. (2024) employ LLMs as knowledge retrievers and separate LMs for generation. SAGE (Chen et al., 2025a) further introduces span-level attention in the generator to focus on a single knowledge record among top-K retrieved candidates, alleviating entity inconsistency.

In summary, while recent pseudo-pipeline methods such as MAKER and SAGE attempt to separate knowledge retrieval and response generation, they still rely on a single model to perform each sub-task. In contrast, our SMARTS framework explicitly coordinates a small model for style-only response generation and a large model for knowledge retrieval and verification, enabling a clear functional separation and enhanced interpretability.

3 Method

3.1 Problem Definition

For the i-th dialogue turn, the dialogue context d_i consists of the full dialogue history and the current user utterance u_i, and is defined as $d_i = [u_1, s_1, \ldots, u_{i-1}, s_{i-1}, u_i]$, where u_j and s_j denote the user utterance and system response at turn j. An external knowledge base K is used to provide knowledge for TOD system to generate informative response. The knowledge base contains N knowledge records, denoted as $K = [e_1, e_2, \ldots, e_N]$. Each knowledge record e_i is a collection of attribute–value pairs and can be represented as $e_i = [(a_i^1, v_i^1), (a_i^2, v_i^2), \ldots, (a_i^j, v_i^j)]$, where a_i^j denotes the j-th attribute and v_i^j denotes its corresponding value. The set of attributes for record e_i is denoted as $A_i = [a_i^1, a_i^2, \ldots, a_i^M]$, where M is the total number of attributes in the record. The system generates a response s_j based on the j-th turn dialogue context d_j and the knowledge base K.

3.2 Model Architecture

The proposed SMARTS consists of a response style generator, a fine-to-coarse knowledge retriever, and a re-thinker. The overall architecture of SMARTS is shown in Figure 2. Firstly, the response style generator according given i-th turn dialogue context d_i to generate the style-only response $\hat{p}_i$. The style-only response is not a completed response,

there are some placeholder attribute slots $A_i^{'}$ in the response. These attribute slots are used as attribute filters for effective knowledge retrieval. That means for each record in the knowledge base, only the attributes corresponding to the slots are retained for subsequent top-1 knowledge retrieval, while the remaining attributes are removed by an attribute filter. Next, the parameter-frozen fine-to-coarse knowledge retriever is utilized to retrieve the most relevant knowledge based on dialogue context d_i and an attribute filtered knowledge base $K^{'}$. The value of the attributes from the retrieved knowledge are filled into the style-only response and form the i-th turn completed response $\hat{s}_i$.

3.3 The Response Style Generator

The response style generator is designed to rapidly generate a style-only response that includes specific attribute slots $A_i^{'}$, which are essential for both fine-to-coarse retrieval and final response generation. Given the i-th dialogue context d_i, an encoder Enc_d transforms the raw dialogue context input into a vector representation h_i^{ctx}. Due to Enc_d consists of a SLM, the context length is limited, here we use a last recent used strategy to truncate the full dialogue context d_i to truncated dialogue context d_i^{trn}, and obtain a corresponding vector representation $h_i^{ctx'}$. The vector $h_i^{ctx'}$ is then passed to a decoder Dec_d with specific adapter layer, which decodes the embedded context into a predicted style-only response $\hat{p}_i$. This response is not fully completed and instead contains attribute slots $A_i^{'}$, placeholders for specific details that will be filled after knowledge retrieval. Formally,

$$\hat{p}_i = LM(d_i^{trn}, \theta_{lm}, \theta_{ap}), \tag{1}$$

where θ_{lm} are the parameters of the SLM used for response style generator, θ_{ap} are parameters of the specific adapter layer.

To comprehensively train the response style generator, we adopt a multi-task training strategy. In addition to generating style-only responses denoted as $\hat{p}_i$, the model is also tasked with predicting the corresponding completed response $\hat{s}_i$. This auxiliary task encourages the model to not only capture the stylistic structure of responses but also develop a better understanding of fine-grained knowledge usage. To facilitate this, we introduce a completed response task adapter layer with parameters θ_{af} into the decoder Dec_d, enhancing the generator's versatility and its ability to generalize across varying dialogue contexts. Formally,

$$\hat{s}_i = LM(d_i^{trn}, K, \theta_{lm}, \theta_{af}). \tag{2}$$

It is worth noting that the completed response generation task is only used in the training phase.

To train these two types of responses, we employ a straightforward yet effective approach using cross-entropy (CE) loss. The specific loss function used in the training process is described in Equation 3.

$$\mathcal{L}_{total} = CE(\hat{p}_i, p_i) + CE(\hat{s}_i, s_i) \tag{3}$$

The p_i is the golden style-only response and s_i is the golden completed response from datasets.

3.4 The Fine-to-Coarse Knowledge Retriever

To enable more precise knowledge retrieval, we propose a fine-to-coarse retrieval mechanism that incorporates attribute-based filtering. Each knowledge record in K is associated with multiple attributes (e.g., "name", "phone", "address", and "type"). However, in most real-world scenarios, many attributes in the retrieved knowledge record are redundant and not ultimately used in the final response. To address this, we extract a subset of relevant attribute slots, denoted as $A_i^{'}$, to retain only the necessary information and discard irrelevant attributes from the knowledge base. The attribute-based filtering process is formally defined in Equation 4.

$$K^{'} = \mathrm{Filter}(K, A_i^{'}) \tag{4}$$

$K^{'}$ is a fine-grained attribute knowledge base obtained by applying a simple filtering function Filter to K, which retains only the attributes specified in $A_i^{'}$ for each knowledge record. Next, the fine-to-coarse knowledge retriever takes both the filtered knowledge base $K^{'}$ and the dialogue context d_i as input. The retriever leverages the LLM's inference capabilities to assess the relevance of each candidate knowledge record, evaluating how well it matches the dialogue context. The used prompt P^r designed as "*Dialogue context: {}. Knowledge records: {}. According to the dialogue context, determine which knowledge records are appropriate. Answering with the name of the knowledge record is fine.*". Through this process, the model directly selects the top-1 most relevant knowledge record for the i-th turn e_i^{top}, ensuring the most accurate and contextually appropriate response is generated. Formally,

$$e_i^{top} = \text{LLM}(d_i, K^{'}, P^r). \quad (5)$$

Given top-1 most relevant knowledge record e_i^{top}, the value of the attributes from the retrieved top-1 relevant knowledge record are filled into the style-only response and form the i-th turn completed response $\hat{s}_i$. Formally,

$$\hat{s}_i = \text{Fill}(e_i^{top}, \hat{p}_i). \quad (6)$$

3.5 The Re-Thinker

The re-thinker is designed to reassess and refine the completed response as by given the dialogue context d_i and retrieved top-1 knowledge record. The re-thinker evaluates this completed response by concatenating it with the dialogue history to assess whether the response is appropriate in the given context. The prompt used for re-think is designed as "*Dialogue context: {}. A response: {}. Do you think the response fits the context? Answer yes or no is enough. Think step by step.*"

If the response is deemed appropriate, this completed response is used as the final completed response. Otherwise, a re-generation prompts P^g is designed to guide the LLM to generate a contextually response. The prompt is "*Dialogue context: {}. Knowledge records: {}. An example of a non-applicable response: {}. You need to respond to the user based on the given dialogue context, the non-applicable response sample, and the knowledge record I have provided to you. So, your response is?*"

This re-thinker ensures that SMARTS output is both contextually knowledge relevant and style-compliant, while also being refined through an additional layer of evaluation to guarantee its overall quality and appropriateness.

3.6 The Inference Phase

In the inference phase in SMARTS, the fine-tuned response style generator, with its compact parameter size, rapidly produces an initial style-only response $\hat{p}_i$. Attribute slots is extracted from $\hat{p}_i$ to form a filtered knowledge base $K^{'}$, excluding irrelevant attributes. The fine-to-coarse knowledge retriever performs top-1 knowledge retrieval using $K^{'}$ and dialogue context, filling attribute values into $\hat{p}_i$ to create a completed response $\hat{s}_i$. The re-thinker evaluates $\hat{s}_i$ for quality and contextual appropriateness. If $\hat{s}_i$ meets quality standards, it serves as the final response; otherwise, a re-generation phase employs an LLM to produce a refined, high-quality response. This mechanism ensures SMARTS balances speed and accuracy. The overall pseudocode of the SMARTS inference phase is presented in Algorithm 1.

Algorithm 1: Pseudo code of SMARTS inference

Input: i-th turn Dialogue context d_i, Knowledge base K

Output: Response from i-th turn dialogue $\hat{s}_i$

$\hat{p}_i, A_i^{'} \leftarrow \text{LM}(d_i^{trn})$ // Obtain a style-only response and attribute slots

$K^{'} \leftarrow \text{Filter}(K, A_i^{'})$ // Obtain a attribute filtered knowledge base according to the attribute slots

$e_i^{top} \leftarrow \text{LLM}(d_i, K^{'}, P^r)$ // Obtain the top-1 knowledge record by a dialogue context and a attribute filtered knowledge base

$\hat{s}_i \leftarrow \text{Fill}(e_i^{top}, \hat{p}_i)$ // Fill the attribute values into the style-only response and obtain a completed response.

$r \leftarrow$ Re-Thinker$(\hat{s}_i, d_i)$ // Determine whether the response is logical based on the dialogue context and return a context-sensitive boolean value r.

if r *is No* **then**

  $\hat{s}_i \leftarrow \text{LLM}(d_i, \hat{s}_i, e_i^{top}, P^g)$ // Re-generate a response as the final completed response.

4 Experiment

4.1 Dataset

The effectiveness of the proposed model was validated using two publicly available task-oriented dialogue datasets: CamRest (Wen et al., 2017), and MultiWOZ 2.1 (MWOZ) (Budzianowski et al., 2018). The Camrest dataset primarily focuses on the restaurant reservation domain and contains 676 multi-turn dialogues. The MWOZ dataset consisted of five domains: train, hotel, restaurant, taxi, and attraction. It is composed of 2,877 multi-turn dialogues. Each dataset was further divided into training, validation, and test sets following previous work (Wan et al., 2023).

4.2 Evaluation Metrics

To evaluate model performance, we use five metrics: BLEU (4-gram), Entity F1, Entity Precision,

Dataset	Method	BLEU	Entity F1	Entity Precision	Entity Recall	Entity Accuracy
CamRest	GraphMemDialog (Wu et al., 2022)	22.30	64.40	-	-	-
	ECO (Huang et al., 2022)	18.42	71.56	-	-	-
	DialoKG (Rony et al., 2022)	23.40	75.60	77.31	73.97	50.32
	MPEToDs (Qin et al., 2023b)	19.30	58.90	-	-	-
	PluDG (Dong and Chen, 2023)	23.00	76.90	77.20	76.61	51.86
	MAKER (Wan et al., 2023)	25.04	73.09	73.89	72.31	46.32
	DF-TOD (Shi et al., 2023)	25.85	72.83	-	-	-
	RGA-TOD (Shen et al., 2023)	25.00	72.09	-	-	-
	Uni-TOD (Ding et al., 2024)	24.70	77.80	78.89	76.78	52.39
	IEM (Chen et al., 2024)	25.99	77.12	77.34	76.91	51.05
	MLTOD (Dong et al., 2024)	25.20	77.49	77.85	77.13	51.79
	SMARTS(Ours)	**27.21**	**79.72**	**79.92**	**79.53**	**54.43**
MWOZ	GraphMemDialog (Wu et al., 2022)	14.90	40.20	-	-	-
	ECO (Huang et al., 2022)	12.61	40.87	-	-	-
	DialoKG (Rony et al., 2022)	12.60	43.50	43.87	43.12	31.47
	MPEToDs (Qin et al., 2023b)	13.60	36.60	-	-	-
	PluDG (Dong and Chen, 2023)	9.20	42.40	43.56	41.32	30.28
	MAKER (Wan et al., 2023)	17.23	53.68	55.17	52.27	37.21
	DF-TOD (Shi et al., 2023)	18.26	52.52	-	-	-
	RGA-TOD (Shen et al., 2023)	16.97	51.99	-	-	-
	Uni-TOD (Ding et al., 2024)	12.30	44.30	46.27	42.49	33.07
	IEM (Chen et al., 2024)	15.18	45.20	47.42	43.18	32.89
	SMARTS(Ours)	**18.82**	**58.55**	**57.48**	**59.65**	**41.07**

Table 1: Performance comparison of SMARTS and baseline models on two public datasets (CamRest and MWOZ). Bold numbers indicate the state-of-the-art results. Numbers with underline indicate the second results.

Entity Recall, and Entity Accuracy. BLEU (Papineni et al., 2002) measures fluency via n-gram overlap between generated and reference text. Entity F1 (Eric et al., 2017), the harmonic mean of precision and recall, assesses accurate entity identification and use in dialogue. Entity Precision evaluates the correctness of generated entities, Entity Recall measures coverage of reference entities, and Entity Accuracy gauges overall entity recognition accuracy.

4.3 Baseline

To objectively evaluate our method, we select a variety of representative baseline models for comparison, categorizing them according to their knowledge retrieval strategies in response generation[1].

Implicit retrieval: These methods integrated the processes of knowledge retrieval and response generation within a single model, including GraphMemDialog (Wu et al., 2022), ECO (Huang et al., 2022), DialoKG (Rony et al., 2022), MPEToDs (Qin et al., 2023b), PluDG (Dong and Chen, 2023), Uni-ToD (Ding et al., 2024) and IEM (Chen et al., 2024).

Explicit retrieval: These methods utilize a knowledge retriever to retrieved the top-K relevant knowledge records, only conduct fine-grained knowledge retrieval during the response generation, including Q-TOD (Tian et al., 2022), MAKER (Wan et al., 2023), DF-TOD (Shi et al., 2023), RGA-TOD (Shen et al., 2023), and MLTOD (Dong et al., 2024).

4.4 Implementation

We use the base version of Flan-T5 (Raffel et al., 2020) as the response style generator. During training, we set the learning rate to 5×10^{-5} and employ the AdamW optimizer. Additionally, we utilize a Llama-3-8B model with frozen parameters as both the fine-to-coarse knowledge retriever and the re-thinker. All experiments are conducted on an NVIDIA RTX 3090 GPU with 24GB of memory.

4.5 Main Results

To comprehensively compare the performance of SMART with several baseline models, we utilize two public datasets and five metrics to evaluate these methods. SMART consistently outperforms baseline models in all metrics across the datasets and achieved state-of-the-art (SOTA). On the CamRest dataset, SMART achieves a BLEU score of 27.21, which represents a 4.7% improvement over the second-best baseline, IEM (25.99). In terms of entity metrics, SMART surpass baselines by 2.5%

[1]The descriptions of these baselines are detailed in the section 2.

Method	BLEU	Entity F1	Entity Precision	Entity Recall	Entity Acc.
Ours	**18.82**	**58.55**	**57.48**	**59.65**	**41.07**
w/o fine-to-coarse knowledge retriever	14.34 (4.48↓)	55.64 (2.91↓)	54.47 (3.01↓)	56.87 (2.78↓)	39.22 (1.85↓)
w/o completed response generation task	16.32 (2.50↓)	56.32 (2.23↓)	55.72 (1.76↓)	56.93 (2.72↓)	39.93 (1.14↓)
w/o re-thinker	17.01 (1.81↓)	57.31 (1.24↓)	56.76 (0.72↓)	57.87 (1.78↓)	40.75 (0.32↓)

Table 2: Model ablation experiments for two different modules on MWOZ dataset.

Retrieval method	Generation method	BLEU ↑	Entity F1 ↑	Time (s) ↓
LLM Coarse-only Retrieval	LLM Generation	7.41	41.39	16.98
LLM Coarse to Fine Retrieval	LLM Generation	9.74	43.09	23.67
LLM Fine to Coarse Retrieval	LLM Generation	10.34	47.05	14.32
LLM Coarse-only Retrieval	LM Generation	14.64	51.95	12.52
LLM Coarse to Fine Retrieval	LM Generation	16.49	53.95	16.38
LM Retrieval	LM Generation	17.23	53.68	**10.89**
LLM Fine to Coarse Retrieval	LM Generation	**18.82**	**58.55**	11.13

Table 3: The impact of large-small language model collaboration framework.

in Entity F1, 1.3% in Entity Precision, 3.1% in Entity Recall, and 3.9% in Entity Accuracy, respectively. On the MWOZ dataset, SMART achieves a BLEU of 18.82, improving by 3.1% over the previous SOTA baseline DF-TOD (18.26). Additionally, SMART outperforms LQ-TOD by 9.1% in Entity F1, 10.1% in Entity recall, and 5.9% in Entity accuracy. compared to MAKER, SMART achieves a 4.2% boost in Entity Precision. These results demonstrate SMART has an advantage in generating fluent, and knowledge-accurate response, highlighting effectiveness of our small-to-large model coordinated retrieval and response framework.

4.6 Ablation study

To explore the contribution of each part from SMART, we conduct the ablation study experiment. Table 2 presents the results of the experiment. When we remove the fine-to-coarse knowledge retriever (*w/o fine-to-coarse knowledge retriever*) and use the coarse-to-fine knowledge retriever from MAKER (Wan et al., 2023) to conduct knowledge retrieval, the performance is depressing. As the same, when we remove the re-thinker (*w/o re-thinker*), SMART also has a performance decline. In addition, removing the completed response generation task (*w/o completed response generation task*) also leads to a noticeable drop across most evaluation metrics, showing that this module is essential for ensuring accurate and coherent response generation. These results indicate that the fine-to-coarse knowledge retriever, the re-thinker, and the completed response generation task all play crucial roles in the overall effectiveness of SMART, as their removal negatively impacts performance across all evaluation metrics.

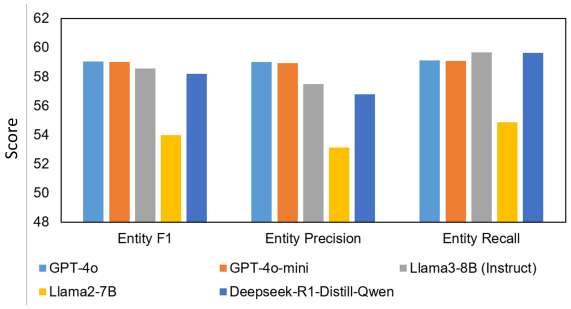


Figure 3: Different LLMs Performance in fine-to-coarse retrieval.

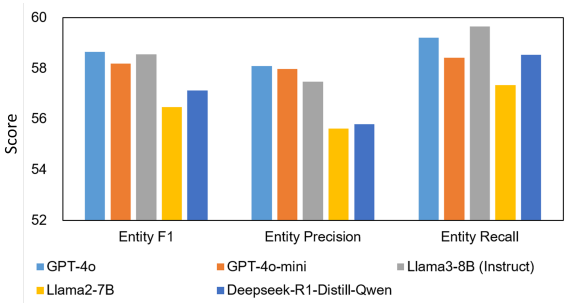


Figure 4: Performance of different LLMs in the Re-thinking.

4.7 Coordinated Strategy of Large-small Language Model Coordinated Framework

In order to explore how to coordinate large language model and small language model to achieve the best performance for TOD, we try several model-coordinated strategies. Table 3 presents an analysis of different model-coordinated strategies between large and small language models within the coordinated framework of SMARTS. The table compares various retrieval methods—LLM

based coarse-only retrieval, LLM based coarse-to-fine retrieval, and LLM based fine-to-coarse retrieval—and their corresponding generation methods, evaluating their impact on two metrics: BLEU, Entity F1. The results reveal that the best-performing configuration is the strategy of LLM based fine to coarse retrieval with LM based generation, achieving the highest BLEU score (18.82) and Entity F1 (58.55). These results suggest that proper strategy between the large and small language models enhances both the quality and efficiency of response generation, with the LLM-based fine-to-coarse retrieval and LM-based generation strategy yielding the most balanced and effective results.

4.8 Impact of Large Language Model on Fine-to-coarse Retrieval and Re-thinking

To investigate how different LLMs affect fine-to-coarse retrieval and re-thinking in the SMARTS framework, we evaluate five LLMs on the MultiWOZ dataset: GPT-4o, GPT-4o-mini, Llama2-7B, Llama3-8B (Dubey et al., 2024), and Deepseek-R1-Distill-Qwen (7B) (Guo et al., 2025). We assess performance using entity-related metrics: Entity F1, Entity Precision, and Entity Recall.

Figures 3 and 4 present the results. GPT-4o and GPT-4o-mini exhibit comparable performance, with GPT-4o-mini slightly outperforming Llama3-8B in Entity Precision. Llama3-8B and Deepseek-R1-Distill-Qwen show similar results across all metrics. In contrast, Llama2-7B performs the weakest, particularly in Entity Precision, where it significantly trails other models.

Analysis reveals that LLMs with stronger instruction-following capabilities (GPT-4o, GPT-4o-mini, Llama3-8B, and Deepseek-R1-Distill-Qwen) achieve similar performance in knowledge retrieval, while Llama2-7B, with weaker instruction-following, underperforms. These findings indicate that instruction-following ability, rather than model size, is critical for effective retrieval and rethinking in the SMARTS framework.

4.9 Performance Comparison on Large-Scale Knowledge Base

Previous methods are typically trained and evaluated on standardized-scale knowledge bases. However, in real-world scenarios, knowledge is retrieved from large-scale knowledge base. To simulate such scenarios, following previous work (Ding et al., 2024; Wan et al., 2023) we aggregate all knowledge records from the databases provided from the CamRest and MWOZ datasets to construct a unified large-scale dataset to evaluate the performance of the methods. The results are shown in Table 4. Our results show that our model outperforms the baselines when using the large-scale knowledge base. It reveals that the performance of existing systems significantly deteriorates when using the large-scale knowledge base compare to the performance in standardized-scale knowledge bases. This decline can be attributed to our model's fine-to-coarse knowledge retrieval strategy and re-think mechanism, which enable it to retrieve more precise knowledge and relevant attributes during response generation.

Method	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	<u>16.67</u>	47.13
MAKER	<u>26.19</u>	<u>72.09</u>	16.25	<u>50.87</u>
Ours	**26.56**	**77.17**	**17.41**	**56.45**

Table 4: The results of SMARTS and baseline models with a large-scale knowledge base on the MWOZ and CamRest datasets respectively.

Model	Naturalness	Correctness	Task Completion
IEM	3.9	3.7	3.8
MAKER	3.8	4.0	3.9
Uni-TOD	3.9	3.8	3.9
SMARTS	**4.1**	**4.4**	**4.3**

Table 5: The result of human evaluation on MWOZ dataset with stateful configuration.

4.10 Human Evaluation

We conduct a human evaluation by distributing an online survey to 50 university volunteers. Our method, SMARTS, is compared with two previous state-of-the-art (SOTA) models, MAKER and Uni-TOD. The evaluation assesses response naturalness, correctness, and task completion using 50 dialogue samples from the MWOZ dataset. The results, presented in Table 5, show that SMARTS consistently achieves the highest scores across all metrics. These findings suggest that SMARTS integrates a more effective mechanism for generating human-like, contextually appropriate, and goal-oriented responses, making it a strong candidate for task-oriented dialogue systems.

5 Conclusion

In this paper, we propose SMARTS, a novel small-to-large model coordinated framework for task-oriented dialogue. SMARTS addresses the task-coupling limitations of traditional TOD systems by introducing a fine-to-coarse knowledge retrieval strategy and a re-think mechanism. These designs enhance retrieval precision, reduce computational complexity, and improves response quality. By coordinating small and large language models, SMARTS effectively balances efficiency and informativeness. Extensive experiments on two public datasets demonstrate its superior knowledge retrieval accuracy and response generation capabilities, achieving significant improvements in Entity F1.

Acknowledgements

The work is supported by grants from National Natural Science Foundation of China (No. 62372189) and Guangdong Special Funding for Science and Technology Innovation Strategy (2024A1111120004).

Limitations

Although SMARTS leverages the strengths of both large and small language models to achieve superior performance, this collaboration also increases the inference time. Additionally, the large language models used in SMARTS can still produce misleading or incorrect information, as they may generate responses based on false or incomplete knowledge. Future work will focus on improving the efficiency of reasoning and addressing these reliability issues.

References

Paweł Budzianowski and Ivan Vulić. 2019. Hello, it's gpt-2-how can i help you? towards the use of pre-trained language models for task-oriented dialogue systems. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 15–22.

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.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. 2019. Skeleton-to-response: Dialogue generation guided by retrieval memory. 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 1219–1228.

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. 2025a. 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, Xuelian Dong, Wenxiu Xie, Ru Peng, Kun Zeng, and Tianyong Hao. 2025b. Llm-enhanced query generation and retrieval preservation for task-oriented dialogue. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14307–14321.

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.

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.

Xuelian Dong, Jiale Chen, Wenxiu Xie, Fu Lee Wang, and Tianyong Hao. 2026. Horizon: A unified framework for phase-wise retrieval-generation optimization in task-oriented dialogue system. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16972–16976. IEEE.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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.

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.

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.

Libo Qin, Wenbo Pan, Qiguang Chen, Lizi Liao, Zhou Yu, Yue Zhang, Wanxiang Che, and Min Li. 2023a. End-to-end task-oriented dialogue: A survey of tasks, methods, and future directions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5925–5941.

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. 2023b. 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.

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.

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.

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.

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.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14230–14238.