LLMs as Collaborator: Demands-Guided Collaborative Retrieval-Augmented Generation for Commonsense Knowledge-Grounded Open-Domain Dialogue Systems

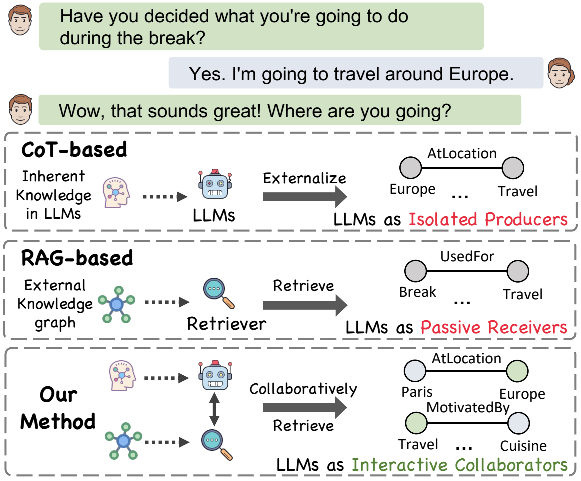
Jiong Yu[1,2], **Sixing Wu**[1,2*], **Jiahao Chen**[1,2], **and Wei Zhou**[1,2]
[1]National Pilot School of Software, Yunnan University, Kunming, China
[2] Engineering Research Center of Cyberspace, Yunnan University, Kunming, China
yujiong@mail.ynu.edu.cn, wusixing@ynu.edu.cn

Abstract

Capturing the unique knowledge demands for each dialogue context plays a crucial role in commonsense knowledge-grounded response generation. However, current *CoT*-based and *RAG*-based methods are still unsatisfactory in the era of LLMs because 1) *CoT* often overestimates the capabilities of LLMs and treats them as isolated knowledge *Producers*; thus, *CoT* only uses the inherent knowledge of LLM itself and then suffers from the hallucination and outdated knowledge, and 2) *RAG* underestimates LLMs because LLMs are the passive *Receivers* that can only use the knowledge retrieved by external retrievers. In contrast, this work regards LLMs as interactive *Collaborators* and proposes a novel *DCRAG*[1] (*Demands-Guided Collaborative RAG*) to leverage the knowledge from both LLMs and the external knowledge graph. Specifically, *DCRAG* designs three *Thought-then-Generate* stages to collaboratively investigate knowledge demands, followed by a *Demands-Guided Knowledge Retrieval* to retrieve external knowledge by interacting with LLMs. Extensive experiments and in-depth analyses on English *DailyDialog* and Chinese *Diamante* datasets proved *DCRAG* can effectively capture knowledge demands and bring higher-quality responses.

1 Introduction

Large Language Models (LLMs) have shown remarkable results in recent studies (Touvron et al., 2023; Zhao et al., 2023). Thanks to the extraordinary conversational abilities of LLMs (Deng et al., 2023), there has been an increasing interest in developing LLM-based open-domain dialogue response systems (Wang et al., 2023b; Chen et al., 2023b). Nonetheless, this development still suffers from many thorny challenges since LLMs are constrained by the *knowledge boundary* (Ren et al., 2023) and lack the *long-tail knowledge* (Kandpal et al., 2023). To mitigate such issues, augmenting the dialogue generation with commonsense knowledge is a promising way, i.e., *Commonsense Knowledge-Grounded Response Generation (CKRG)*. Commonsense knowledge can significantly deepen the understanding of LLMs about the real world (Guan et al., 2024).


Have you decided what you're going to do during the break?

Yes. I'm going to travel around Europe.

Wow, that sounds great! Where are you going?

Figure 1: A comparison between previous LLM-based CKRG paradigm and our method. We propose to regard LLMs as *Collaborators* to leverage the knowledge from both LLMs and the external knowledge graph.

Although massive commonsense facts are in the world, only a few highly relevant facts are indeed demanded by each specific dialogue context. Thus, how to seek commonsense facts to meet the unique *Knowledge Demands* of each dialogue context is fundamental in CKRG. Prior works can be either *CoT-based* or *RAG-based*. On the one hand, benefiting from the scale of params and training corpora, a large-scale LLM itself can be regarded as an inherent knowledge base (Chen et al., 2023a; Bian et al., 2024). Thus, *CoT-based* works (Zhou et al., 2022b; Liu et al., 2022; Chae et al., 2023) externalize the implicit knowledge from the backbone LLMs in ad-

[*]The corresponding author.

[1]Our codes and dataset are released at https://github.com/Y-NLP/Chatbots/tree/main/EMNLP2024_DCRAG.

Findings of the Association for Computational Linguistics: EACL 2024, pages 13586–13612
November 12-16, 2024 ©2024 Association for Computational Linguistics

vance to assist the following response generation in the manner of Chain-of-Thought (Wei et al., 2022). On the other hand, external commonsense knowledge graphs, like ConceptNet (Speer et al., 2017) and Atomic (Sap et al., 2019), consist of various high-quality human-collected facts. Thus, *RAG-based* (Retrieval-Augmented Generation) methods (Zhao et al., 2024) explicitly retrieve facts from external knowledge bases to augment the response generation (Wu et al., 2020; Gao et al., 2022).

Nonetheless, both mentioned paradigms are still unsatisfactory in capturing the knowledge demands in the era of LLMs because inappropriate roles are assigned to LLMs. In the *CoT-based* paradigm, LLMs act as isolated *Producers*, and the knowledge is totally externalized from the LLM's internal parameters without external interaction. However, this paradigm heavily relies on the LLM scale and overestimates its capabilities; thus, it is inevitably limited by hallucination (Ji et al., 2023) and outdated (Wang et al., 2023c) knowledge. Conversely, *RAG-based* paradigm underestimates the capabilities of LLMs by only treating them as passive *Receivers*, which can only use the knowledge retrieved by an independent external retriever. However, the notable cognition gap between the retriever and LLMs creates an information bottleneck. The independent retriever may miss implicit knowledge demands like ambiguous intentions and possible involved topics, and LLMs lack the chance to help the retrieval in spite of their great power.

This work regards the LLMs as interactive *Collaborators* and presents a novel *DCRAG* (*Demands-Guided Collaborative RAG*). As illustrated in Figure 1, compared to prior paradigms, *DCRAG* not only uses the inherent knowledge from the LLM itself but also leverages the great power of LLMs to collaboratively retrieve external commonsense knowledge. In detail, *DCRAG* thinks knowledge demands are derived from the queries and topics, where queries are knowledge explicitly or implicitly mentioned by the dialogue history, and topics are the possible involved knowledge in the subsequent response. Thus, *DCRAG* designs three *Thought-then-Generate* stages: *Query Production*, *Topic Planning*, and *Cross Revision* to collaboratively investigate knowledge demands with LLMs. Then, based on the identified knowledge demands, *DCRAG* can more effectively retrieve the commonsense knowledge via *Demands-Guided Knowledge Retrieval* to ground the following response generation by further interacting with LLMs.

We have conducted experiments on English *DailyDialog* (Li et al., 2017) and Chinese *Diamante* (Lu et al., 2023) datasets. Besides the existing reference-free metrics, we propose *CDP* and *CDF* to measure the usage of commonsense knowledge and *GPT-4 Evaluation* metrics to measure the various aspects of responses. Extensive experiments and in-depth analyses demonstrated *DCRAG* can effectively capture knowledge demands and bring higher-quality responses compared to baselines.

Our contribution is three-fold: 1) We propose to regard LLMs as interactive *Collaborators* to better capture the unique knowledge demands of each dialogue context without overestimating or underestimating the capabilities of LLMs; 2) We propose a novel *DCRAG* to collaboratively identify knowledge demands and leverage the knowledge from both LLMs and the external knowledge graph; 3) Extensive experiments and in-depth analyses confirmed the effectiveness of our *DCRAG*.

2 Related Work

Knowledge-Grounded Response Generation (KRG) Recently, KRG has been extensively studied to alleviate the issue of generic or hallucinated responses (Li et al., 2016; Roller et al., 2021; Shuster et al., 2021), which frequently appears in the traditional response generation (RG) systems. KRG can generate more diverse responses grounded on additional knowledge, such as Wikipedia (Dinan et al., 2019; Kim et al., 2020), Internet (Shuster et al., 2022; Wang et al., 2023a), Commonsense (Wu et al., 2020; Tang et al., 2023), and some others (Li et al., 2022; Ni et al., 2023). This work mainly focuses on the KRG with commonsense, which can facilitate the models to grasp factual knowledge about the real world.

Knowledge Seeking in CKRG It is crucial since it can significantly affect the subsequent response. Previous CKRG works externalize knowledge from the LLMs (Zhou et al., 2022b; Liu et al., 2022; Chae et al., 2023), or retrieve external knowledge by independent retriever (Zhou et al., 2018; Wu et al., 2020; Zhou et al., 2021; Gao et al., 2022). However, both of them are unsatisfactory in capturing the knowledge demands since they either are limited by hallucination (Ji et al., 2023) and outdated (Wang et al., 2023c) knowledge or encounter an information bottleneck due to the cognition gap between retriever and LLMs. Thus, this work regards LLMs as *Collaborators* to address this issue.

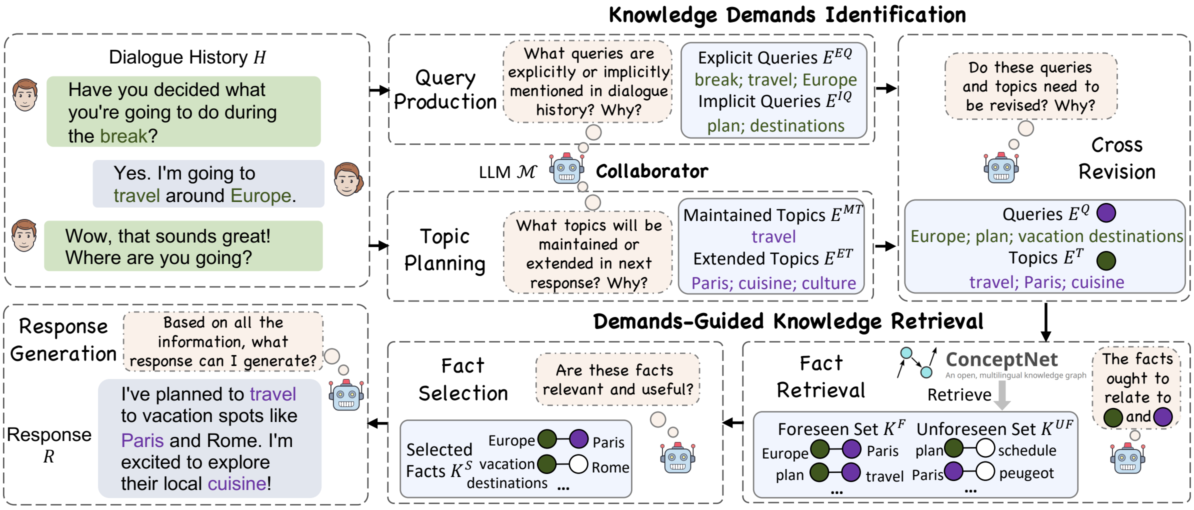

Knowledge Demands Identification
Dialogue History H
Have you decided what you're going to do during the break?
Yes. I'm going to travel around Europe.
Wow, that sounds great! Where are you going?
Query Production
What queries are explicitly or implicitly mentioned in dialogue history? Why?
Explicit Queries E^{EQ}
break; travel; Europe
Implicit Queries E^{IQ}
plan; destinations
LLM $\mathcal{M}$
Collaborator
Topic Planning
What topics will be maintained or extended in next response? Why?
Maintained Topics E^{MT}
travel
Extended Topics E^{ET}
Paris; cuisine; culture
Do these queries and topics need to be revised? Why?
Cross Revision
Queries E^Q
Europe; plan; vacation destinations
Topics E^T
travel; Paris; cuisine
Demands-Guided Knowledge Retrieval
Response Generation
Based on all the information, what response can I generate?
Response R
I've planned to travel to vacation spots like Paris and Rome. I'm excited to explore their local cuisine!
Fact Selection
Are these facts relevant and useful?
Selected Facts K^S
Europe — Paris
vacation destinations — Rome
...
Fact Retrieval
ConceptNet
An open, multilingual knowledge graph
Retrieve
The facts ought to relate to and
Foreseen Set K^F
Europe — Paris
plan — travel
...
Unforeseen Set K^{UF}
plan — schedule
Paris — peugeot
...

Figure 2: An illustration of *DCRAG*. It regards LLM as an interactive *Collaborator* to 1) collaboratively identify the knowledge demands via *Query Production*, *Topic Planning*, and *Cross Revision* and 2) retrieve external knowledge via *Fact Retrieval* and *Fact Selection*. Then, augmenting *Response Generation* with retrieved knowledge.

Queries and Topics In this work, both queries and topics are used to describe knowledge demands in dialogue. Such views are motivated by the current Query Production (Komeili et al., 2022; Shuster et al., 2022; Reddy et al., 2023; Wang et al., 2023a) and the Topic-Guided Response Generation (Tang et al., 2019; Zhong et al., 2021; Tan et al., 2023). The former aims to generate search queries to seek knowledge from external knowledge sources, and the latter predicts the topic to guide knowledge retrieval and response generation.

3 Methodology

3.1 Overview

As illustrated in Figure 2, this work proposes a novel LLM-based *DCRAG*, which regards LLM as *Collaborator* in capturing the knowledge demands and seeking external commonsense knowledge. Given a conversational corpus $\mathcal{D} = \{(H, R)\}^{|\mathcal{D}|}$, where H/R is the dialogue history/response, and an external commonsense knowledge graph $\mathcal{G} = \{k\}^{|\mathcal{G}|}$, where each $k = (e_h, r, e_t)$ is a fact triplet. This work first collaboratively identifies Queries E^Q and Topics E^T via *Knowledge Demands Identification*. Then, based on the E^Q and E^T, it retrieves external knowledge K from $\mathcal{G}$ via *Demands-Guided Knowledge Retrieval*. Finally, the response generation can be defined as $P(R|H, K)$.

3.2 Knowledge Demands Identification

Identifying the knowledge demands is crucial because it affects what kind of commonsense knowledge facts can be retrieved to enhance the following response generation. Prior RAG-based works (Wu et al., 2020; Gao et al., 2022) directly regard all entities that appeared in the dialogue history H as the queries to retrieve knowledge. This paradigm faces three drawbacks: 1) it can not capture the implicit knowledge demands that are not directly mentioned by the dialogue history; 2) not all mentioned entities can reflect the current demands because some of them are irrelevant or noise; 3) it ignores the fact that the topics of the next response may not be fully involved by the current dialogue history.

Differently, we describe knowledge demands using both queries reflecting the intentions and topics possibly involved. Then, we propose *Thought-then-Generate* to collaboratively identify them with three stages: *Query Production*, *Topic Prediction*, and *Cross Revision*.

3.2.1 Query Production

It aims to identify entities that can reflect the explicit and implicit intentions of the current dialogue history H. To investigate both the explicit and implicit demands in H, we design a prompt to instruct the backbone LLM $\mathcal{M}$ to 1) selectively *EXTRACT* the relevant entities E^{EQ} from H and 2) *INFER* the possible entities E^{IQ} implicitly referred by H.

To enable a deeper and more interpretable reasoning, we propose a *Thought-then-Generate* paradigm to generate such queries by collaboratively interacting with the backbone LLM $\mathcal{M}$ in a two-step Chain-of-Thought (CoT) manner:

$$T^{EQ}, T^{IQ} \leftarrow P^{QP}_{\mathcal{M}}(H) \qquad (1)$$

$$E^{EQ}, E^{IQ} \leftarrow P_{\mathcal{M}}^{QP}(H, T^{EQ}, T^{IQ}) \tag{2}$$

where the corresponding prompt $P_{\mathcal{M}}^{QP}$[2] and two equations can be explained as follow:

- *Thought step:* Eq 1 asks $\mathcal{M}$ to think about what entities should be explicitly extracted (T^{EQ}) or implicitly inferred (T^{IQ}) and write the corresponding thoughts as the plain text.
- *Generate step:* Based on the thoughts T^{EQ} and T^{IQ}, Eq 2 then asks $\mathcal{M}$ to selectively extract the explicit queries (i.e., E^{EQ}) and infers the implicit queries (i.e., E^{IQ}) from H.

3.2.2 Topic Planning

Solely investigating the queries from the dialogue history H is not enough to capture the knowledge demands since the topic of the subsequent response may be extended. Thus, it is not trivial to proactively plan the possible topics that can be involved in the subsequent response when identifying the knowledge demands. In detail, we assume the subsequent dialogue response may either maintain the existing topics or extend to new topics:

1. *Maintained Topics E^{MT}:* Some entities explicitly or implicitly mentioned in the dialogue history may be maintained in the subsequent response to ensure the conversation's coherence and engagingness.
2. *Extended Topics E^{ET}:* Entities that are not in the dialogue history H, but may be involved by the subsequent response.

then, similar to the prior stage, we use a prompt $P_{\mathcal{M}}^{TP}$ to ask the backbone $\mathcal{M}$ to collaboratively plan topics in the *Thought-then-Generate* way:

$$T^{MT}, T^{ET} \leftarrow P_{\mathcal{M}}^{TP}(H) \tag{3}$$

$$E^{MT}, E^{ET} \leftarrow P_{\mathcal{M}}^{TP}(H, T^{MT}, T^{ET}) \tag{4}$$

3.2.3 Cross Revision

The former two stages identify the knowledge demands individually to avoid cross-interference. However, we also believe the interaction between them is necessary. Thus, this stage asks the backbone $\mathcal{M}$ to revise the previously identified queries and topics collaboratively based on their interrelations. In detail, we still use a prompt $P_{\mathcal{M}}^{CR}$ to perform this job in the *Thought-then-Generate* way:

$$\hat{T} \leftarrow P_{\mathcal{M}}^{CR}(H, E, T) \tag{5}$$

$$\hat{E} \leftarrow P_{\mathcal{M}}^{CR}(H, E, T, \hat{T}) \tag{6}$$

where the $E = \{E^{EQ}, E^{IQ}, E^{MT}, E^{ET}\}$ and $T = \{T^{EQ}, T^{IQ}, T^{MT}, T^{ET}\}$ are the collections of entities and thoughts respectively, both of which are the outputs of *Query Production* and *Topic Planning*, $\hat{T} = \{T^Q, T^T\}$ and $\hat{E} = \{E^Q, E^T\}$ is the revision thoughts, and revised queries and topics. Then, the collaboratively identified queries E^Q and topics E^T are used to describe the comprehensive knowledge demands for the given dialogue context.

3.3 Demands-Guided Knowledge Retrieval

Once the knowledge demands have been identified, the next step is to retrieve external commonsense knowledge facts to meet these knowledge demands by further interacting with LLM.

Fact Retrieval With the guidance of the identified knowledge demands (i.e., Queries E^Q and Topics E^T), we first retrieve two kinds of facts from the external knowledge graph $\mathcal{G} = \{k_j = (e_h, r, e_t)\}$:

- *Foreseen Set:* It includes facts that are fully planned by the knowledge demands. For each fact $k = (e_h, r, e_t) \in \mathcal{G}$, if $e_h \in E^Q$ and $e_t \in E^T$, or, if $e_h \in E^T$ and $e_t \in E^Q$, then k is added to the foreseen set.
- *Unforeseen Set:* We also retrieve facts that are partially planned because the size of foreseen facts is always limited. For each fact $k = (e_h, r, e_t) \in \mathcal{G}$, if e_h or $e_t \in E^Q \cup E^T$, then k is added to the unforeseen set.

then, for both sets, we use a dense retriever[3] to select top-50 relevant facts as the final foreseen set K^F and unforeseen set K^{UF}.

Fact Selection To further interact with LLM by effectively eliminating irrelevant facts in K^F and K^{UF}, we turn to use the prompt $P_{\mathcal{M}}^{FS}$ to ask $\mathcal{M}$ to pick up foreseen set K^F and unforeseen set K^{UF} facts by considering the relevance and usefulness:

$$K_S \leftarrow P_{\mathcal{M}}^{FS}(H, K^F) + P_{\mathcal{M}}^{FS}(H, K^F) \tag{7}$$

where $P_{\mathcal{M}}^{FS}(H, K^F)$ first orderly selects up to 20 facts from the foreseen set K^F as K_S; then, if the size of $|K_S| < 20$, $P_{\mathcal{M}}^{FS}(H, K^F)$ continues to orderly select facts from K^{UF} until $|K_S| = 20$.

[2] All prompts of DCRAG are reported in Appendix E.2.

[3] This dense retriever *Ranker* is implemented based on *text-embedding-3-small*, an embedding model. It will rank the facts based on the embedding cosine similarity between each fact triplet and dialogue history.

3.4 Response Generation

Finally, we can instruct the backbone LLM $\mathcal{M}$ to generate the subsequent dialogue response:

$$R \leftarrow P_{\mathcal{M}}^{RG}(H, K^S, I) \quad (8)$$

where H is the dialogue history, K^S is the selected facts, and I is the outputs of previous entities and thoughts in *Cross Revision* to help the backbone model understand why K^S can be retrieved.

4 Experiments

4.1 Experiment Settings

4.1.1 Dataset

Dialogues Our experiments use an English *DailyDialog* dataset (Li et al., 2017) and a Chinese *Diamante* dataset (Lu et al., 2023). Both of them are human-to-human multi-turn conversation corpus in the open-domain settings. The English *DailyDialog* dataset contains 12,539 dialogues with 94,400 utterances, and the Chinese *Diamante* dataset contains 6,838 dialogues with 98,115 utterances. Considering the high cost of LLMs, we randomly sample 300 instances as the test sets from the original part of *DailyDialog* and *Diamante*, respectively.

Knowledge We employ a widely-used multilingual commonsense knowledge graph, ConceptNet (Speer et al., 2017). It contains 1.17M entities, 3.28M facts, and 47 relations in English; 0.13M entities, 0.37M facts, and 25 relations in Chinese.

4.1.2 Models

We tested three types of models. Except *Cosmo*, all models are built upon the `gpt-3.5-turbo-1106`. The implementation details are reported in Appendix A.1. The first **RG** type does not consider retrieving or externalizing the knowledge in the inference. *1) Vanilla* directly adopts a prompt to drive the backbone LLM to generate the response. *2) Cosmo*[4] is further trained on a commonsense knowledge-enhanced dataset SODA (Kim et al., 2023). The next is **CoT-based CKRG**. *3) MSDP* (Liu et al., 2022) uses a multi-stage prompting framework to externalize the implicit knowledge before generating the response. The last is **RAG-based CKRG**, which explicitly retrieves knowledge from the external commonsense knowledge graph. Depending on how to obtain the queries, we consider three methods: *4) Traditional RAG (T-RAG):* Following (Wu et al., 2020; Gao et al., 2022), it regards all the entities that appeared in the dialogue history as queries. *5) Query Production (QP-RAG):* Inspired by the current query production works (Komeili et al., 2022; Zhou et al., 2022a; Wang et al., 2023a), it builds an independent learnable generation network to produce the queries. *6) SCG-QP:* (Reddy et al., 2023) generates queries by leveraging the Cosmo model (Kim et al., 2023) to establish connections related to the conversation topic. For all RAG-based models, after obtaining the queries, we retrieve the facts that head (or tail) entity in the queries and then employ the dense retriever (see Sec 3.3) to re-rank facts and select the top 20 as the retrieved knowledge. Finally, we instruct the backbone LLM to generate the responses.

4.1.3 Automatic Metrics

Since most models are built on LLMs without being fine-tuned on the specific datasets, using the reference-based metrics to evaluate responses is not suitable (Li et al., 2024). Hence, we use the *Reference-free* and *GPT-4 Evaluation* metrics.

Reference-free We use *Distinct-2* (**DI-2**) to measure the diversity of the responses (Li et al., 2016). Then, to measure the usage of commonsense knowledge in responses, we propose **CDP** and **CDF**.

1. **CDP** *(Commonsense Dialogue Proportion)* shows the proportion of knowledge-grounded responses. We use *Hard Matching* (Zhou et al., 2022b) to extract fact tuples $\{t\}$ between the dialogue history and the generated response, and then:

$$\textbf{CDP} = \frac{1}{n}\sum_{1}^{n} \mathbb{I}(\{t\}), \mathbb{I} = 1 \text{ if } \{t\} \neq \emptyset \textit{ else } 0$$

where n is the test set size, $\mathbb{I} = 1$ if the response is grounded by at least one commonsense fact.

2. **CDF** *(Commonsense Dialogue Feature)* considers the abundance of informative fact tuples that appear in the generated response using the IDF value (Ramos et al., 2003):

$$\textbf{CDF} = \frac{1}{n}\textstyle\sum_{1}^{n}\sum_{1}^{|\{t\}|}\text{IDF}(t)$$

GPT-4 Evaluation Recent works like GPTScore (Fu et al., 2023) and G-EVAL (Liu et al., 2023b) have shown the effectiveness of LLMs evaluation in the text generation task. Thus, we employ **MEEP** (Ferron et al., 2023) to measure the engagingness of generated responses based on GPT-4

[4] 3B, `cosmo-xl`.

Pattern	**Method**	*PPL*	*Reference-free*			*GPT-4 Evaluation*			**GeoMean**
			DI-2	**CDP**	**CDF**	**MEEP**	**Infor.**	**Overall**	
English DailyDialog dataset (Li et al., 2017)									
RG	*Vanilla*	26.48	57.46	68.67	22.77	61.40	59.50	86.42	55.23
	Cosmo	20.40	53.00	60.00	14.44	47.10	55.25	76.92	45.77
CoT-based	*MSDP*	35.48	**63.49**	74.33	38.83	48.53	62.17	83.17	59.86
RAG-based	*T-RAG*	27.14	59.27	74.00	27.24	61.38	61.75	86.67	58.30
	QP-RAG	25.55	57.14	75.67	27.52	58.13	59.75	85.42	57.27
	SCG-QP	21.24	57.10	82.00	33.68	59.18	61.83	86.92	60.73
DCRAG (Ours)		29.01	59.70	**85.67**	**41.60**	**71.48**	**72.92**	**92.75**	**68.45**
Chinese Diamante dataset (Lu et al., 2023)									
RG	*Vanilla*	35.06	70.91	45.33	8.63	69.67	60.08	81.67	46.00
CoT-based	*MSDP*	27.58	77.80*	50.00	10.93	70.50	59.42	82.50	49.49
RAG-based	*T-RAG*	36.62	77.17	52.33	11.47	66.07	58.92	81.92	49.53
	QP-RAG	27.94	**78.51**	52.00	11.27*	68.73	59.75	82.58	49.99
DCRAG (Ours)		18.56	77.57	**58.67**	**14.51**	**80.64**	**68.83**	**88.33**	**56.45**

Table 1: Main Results. **GeoMean** is the geomean score. Although *Perplexity (PPL)* is reported, we do not compare this metric (see Appendix A.2). Expect *, *DCRAG* shows significance ($p < 0.05$, Wilcoxon Signed-Rank Test).

(gpt-4-1106-perview) (OpenAI, 2023). Meanwhile, we also instruct GPT-4 to score the generated responses like humans in the aspects of *Informativeness* (**Infor.**) and *Overall Score* (**Overall**). The used prompts are reported in Appendix E.3.

4.2 Main Results

Automatic Evaluation As shown in Table 1, *DCRAG* shows notable superiority on both datasets, proving it can better capture the knowledge demands and then generate higher-quality responses. For *RG* methods, without the help of commonsense knowledge, *Vanilla* generates the least commonsense-grounded responses (due to poor CDP and CDF). *Cosmo* has the worst performance since the used small-scale backbone model (3B params only). As a *CoT-based* method, *MSDP* is notably better than *RG* methods, especially in terms of DI-2. It proves that grounding the response generation on the externalized implicit knowledge is beneficial. In the *RAG-based* paradigm, *T-RAG* achieves comparable overall performance with *MSDP*. Such results demonstrate either retrieving external knowledge or externalizing implicit knowledge in LLMs can meet the knowledge demands to a certain extent. After asking LLM to investigate partial knowledge demands (i.e., queries) in dialogue history, *QP-RAG* and *SCG-QP* show visible improvements. It illustrates the information bottleneck in *T-RAG* can be broken by interacting with LLM. By continuously interacting with LLM by 1) investigating deeper knowledge demands (such as possible topics) and 2) engaging in external knowledge retrieval, *DCRAG* outperforms all baselines by notable margins. Such results confirm the effectiveness of *Collaborator* since it can jointly leverage the strengths of LLMs and external knowledge.

Winner	**Natur.**	**Coher.**	**Engag.**	**Infor.**
Vanilla	35%	36%	21%	10%
DCRAG	**60%**	**61%**	**75%**	**88%**
MSDP	29%	33%	22%	17%
DCRAG	**68%**	**66%**	**77%**	**81%**
T-RAG	32%	36%	20%	12%
DCRAG	**63%**	**62%**	**75%**	**86%**
QP-RAG	35%	35%	19%	17%
DCRAG	**62%**	**63%**	**77%**	**80%**

Table 2: Human Evaluation Results. We report the proportion of each model winning. All differences are significant ($p < 0.05$, Wilcoxon Signed-Rank Test). The average Cohen's kappa among annotators is 0.733.

Pair-wise Human Evaluation We randomly sampled 100 test cases from the Chinese *Diamante* dataset and asked three well-educated native-speaker annotators to select the better response between the two candidates in terms of Naturalness (*Natur.*), Coherence (*Coher.*), Engagingness (*Engag.*), and Informativeness (*Infor.*)[5]. We allowed 'Tie' if the two candidates were comparable. As shown in Table 2, the responses generated by our

[5] See the explanations of criteria in Appendix A.3.

Method	CDP	CDF	MEEP	Overall	GeoMean
DCRAG (Full)	**85.67**	**41.60**	**71.48**	**92.75**	**68.45**
-w/o FS	82.67	38.83	66.25	91.33	65.39
-w/o FR+FS	82.67	34.21	66.88	92.00	64.38
-w/o QP+CR	82.33	36.91	66.50	91.67	64.50
-w/o TP+CR	82.33	36.28	62.00	88.42	63.00
-w/o CR	81.33	36.37	66.24	90.25	64.28
-w/o TtG	76.33	32.18	59.33	86.33	60.58

Table 3: Ablation Study on *DailyDialog*. Here, we remove modules in *DCRAG*, including *Query Production (QP)*, *Topic Planning (TP)*, *Cross Revision (CR)*, *Fact Retrieval (FR)*, and *Fact Selection (FS)*, and a reasoning way of *Thought-then-Generation (TtG)*. *DCRAG* shows significance ($p < 0.01$, Wilcoxon Signed-Rank Test).

DCRAG can better align with the human preferences in all terms, especially the *Engag.* and *Infor.*. We think this is because *DCRAG* can simultaneously identify knowledge demands (i.e., queries and topics) from both dialogue history and subsequent response, which effectively helps the backbone model to jointly maintain existing information and extend new topics in generating responses.

4.3 In-Depth Analysis

Ablation Study To decompose the contribution of *DCRAG*, we conduct an ablation study by removing each module and a reasoning way from full *DCRAG* in Table 3. **First**, we verify the effectiveness of *Demand-Guided Knowledge Retrieval*: *1) -w/o FS* generate responses directly based on the retrieved external facts; the decreased performance highlights the necessity of interacting with LLM to filter out irrelevant facts. *2) -w/o FR+FS* removes entire external retrieval, where *DCRAG* degenerated as an CoT-based method. Although it still outperforms *MSDP* (see Table 1), there remains a significant gap compared to the full *DCRAG*. **Next**, we study the effectiveness of *Knowledge Demands Identification*: *3) -w/o QP+CR* and *4) -w/o TP+CR* only rely on the partial knowledge demands (i.e., queries or topics) to retrieve facts. The performances are all dropped, proving it is accurate to use both of them to describe the knowledge demands. The next *5) -w/o CR* uses all but preliminary queries and topics to retrieve facts, also resulting in poorer performance. We also observed that a decreased performance from *3) -w/o QP+CR* to *5) -w/o CR*. We own that, without the aid of topics, only using queries will retrieve more relevant but useless facts. **Last**, *6) w/o TtG* directly generates entities without *Thought-then-Generation* reasoning, and the sharply declined results suggest it is crucial for enhancing the reasoning performance of LLM.

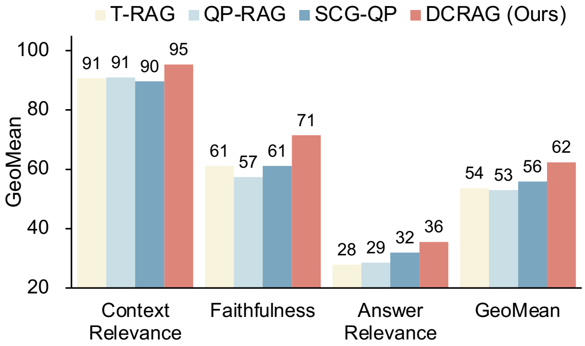


Figure 3: RAG Pipeline Evaluation Results on *DailyDialog* using RAGAs (Es et al., 2024).

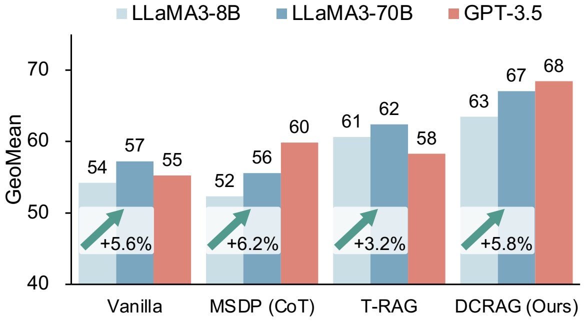


Figure 4: LLM Sensitivity Analysis on *DailyDialog*.

RAG Pipeline Evaluation Prior experiments only focus on the quality of responses, lacking an evaluation to investigate the performance of middle component (i.e., retrieved knowledge) and the entire pipeline. Thus, we employ RAGAs (Es et al., 2024) to evaluate our *DCRAG* and other RAG-based baselines. RAGAs is a framework using reference-free metrics[6] to evaluate RAG pipelines, which consist of Dialogue History (Query), Knowledge (Context), and Response (Answer). As shown in Figure 3, we derive the following findings: 1) According to *Context Relevance* and *Faithfulness*, *DCRAG* can effectively capture knowledge demands since it can retrieve more relevant and useful knowledge facts; 2) *DCRAG* has the ability to generate the response that better aligns with the dialogue (see *Answer Relevance*); 3) The entire pipeline of *DCRAG* is more coordinate and effective due to excellent performance across all metrics.

LLM Sensitivity *DCRAG* can efficiently leverage the knowledge from both the LLM itself and

[6] Appendix A.4 reports the details of RAGAs metrics.

Dialogue History
A: What a confidence! I always watch a lot of movies, too.
B: Oh really, Frank? What kind of movies do you like?
A: Actually, I watch whichever movie there is, be it a comedy, a sci-fi, or a suspense movie. How about you?
B: Art films are my favorite, but thrillers are cool, too.
A: Really impressive. These two genres are different.

Vanilla
R: Yeah, they are quite different, but both offer unique experiences, don't you think?

MSDP
K: (art film, IsA, movie); (thriller, IsA, movie); (genre, Synonym, type); ...
R: Art films are my favorite, but thrillers are cool too. It's impressive how different these two genres are.

T-RAG
K: (frank, RelatedTo, seriously truthful); (sealed room, IsA, film); (frank, RelatedTo, being honest); ...
R: Yeah, they are quite different, but each has its own charm, don't you think?

DCRAG
E^Q: movies; art films; thrillers; reasons behind favorite movie choices; specific interests in movie genres
E^T: movies; suspense movie; art films; thrillers; genres; participants' preferences; differences between movie genres; specific examples of art films; specific examples of thrillers; reasons behind favorite movie choices
K: (movie, RelatedTo, film); (suspense movie, IsA, movie); (art film, RelatedTo, artistic); ...
R: Yes, they are indeed quite different. I love the depth and artistic expression in art films, but the adrenaline rush from thrillers is also exhilarating.

Table 4: Case Study. E^Q, E^T, K, and R denote queries, topics, externalized/retrieved knowledge, and response, respectively. We reported more cases in Appendix C.

the external knowledge graph. To validate its superiority, besides the tested GPT-3.5, we also test *DCRAG* and three baselines on LLaMA3-8B and LLaMA3-70B (Touvron et al., 2023). As illustrated in Figure 4, *DCRAG* meets the expectation indeed since it not only shows superior performance on the small scale LLM but also can achieve notable performance gains from scaling up params (such as from LLaMA3 8B to 70B). In contrast, *MSDP* exhibits the poorest performance with the smallest LLM (GeoMean is only 52 in LLaMA3-8B). It illustrates the limitation that isolated *Producers* heavily relies on the LLM scale. Meanwhile, *T-RAG* underestimates the capabilities of LLMs to treat them as passive *Receivers*, thereby it is hard to benefit from the LLM scale (the percentage increase is only 3.2% from LLaMA3 8B to 70B).

Case Study Table 4 reports one case in *DailyDialog*. It can be seen that *Vanilla* did not perform well since the generated response is generic. The externalized knowledge of *MSDP* fails to meet knowledge demands due to low novelty, and the following response can only repeat the existing information. *T-RAG* retrieves irrelevant facts but has a more acceptable response. Our *DCRAG* first identifies queries and topics to accurately describe the knowledge demands; then, the retrieved facts can better meet the demands. Thus, the response of *DCRAG* is notably knowledgeable and high-quality.

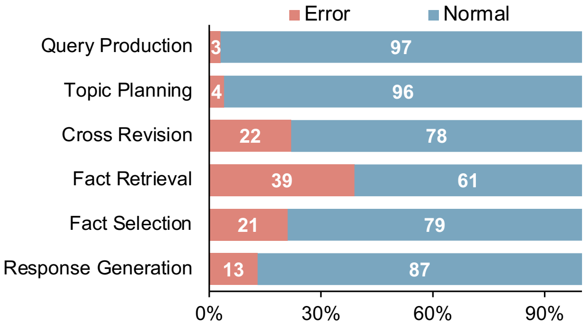


Figure 5: Error Analysis Results. See examples and corresponding analyses for each error type in Appendix D.

Error Analysis To gain a comprehensive understanding of our *DCRAG*, we randomly select 100 instances from *DailyDialog* to manually inspect the errors of each module. The aggregated statistical findings are detailed in Figure 5. We find that Cross Revision, Fact Retrieval, and Fact Selection have more notable error rates. Compared to other modules, we can attribute the reason as the reasoning requirements of such modules are more complex and challenging. Nonetheless, the benefits of integrating them can surpass the subsequent errors. For example, previous ablated results (see Table 3) have proved such modules are crucial for ensuring the quality of queries and topics, and meeting the knowledge demands by external knowledge.

5 Conclusion

Previous LLMs-based CKRG works fail to adequately capture the knowledge demands, as they inappropriately treat LLMs as isolated *Producers* or passive *Receivers*. Differently, this work regards LLMs as interactive *Collaborators* and presents a novel *DCRAG*. It first employs three *Thought-then-Generate* stages to collaboratively identify knowledge demands; then performs *Demands-Guided Knowledge Retrieval* to retrieve external knowledge. Extensive experiments and in-depth analyses have verified *DCRAG* can better capture knowledge demands and generate higher-quality responses.

Acknowledgement

This work is mainly supported by Yunnan Fundamental Research Projects under Grant No.202401CF070170, in part by the Yunnan Province Wu Zhonghai Expert Workstation under Grant No.202305AF150045, and also in part by the National Natural Science Foundation of China under Grant 62162067 and 62101480.

Limitations

The known limitations of our approach can summarized as follows:

- **Inference Overhead** The proposed *DCRAG* is built up on the LLMs and involves six inference steps (Query Production, Topic Planning, Cross Revision, Fact Selection for the foreseen set, Fact Selection for the unforeseen set, and Response Generation) to generate a response for a given dialogue. Like other works in the Chain-of-Thought paradigm, our approach incurs a high inference cost but has to make a trade-off between cost and benefit. In further work, we will investigate how to dynamically adjust the inference steps and objects for specific dialogue scenarios in order to reduce the inference overhead.
- **Requirements for Reasoning Ability of LLMs** *DCRAG* requires the backbone LLMs to accurately identify the knowledge demands, select useful facts, and then generate a response, necessitating a certain level of reasoning ability. We test *DCRAG* on a small LLaMA3-8B and the results are satisfactory (see the LLM Sensitivity experiment in Sec 4.3), while performance on smaller LLMs like Gemma-2B remains uncertain, where this LLM has notably weaker reasoning abilities. We plan to conduct more validation experiments and try to enhance our *DCRAG* performance on smaller LLMs with technologies like knowledge distillation in future work.

Ethical Considerations

This work studies commonsense knowledge-grounded response generation (CKRG), which is a commonly and extensively researched task. We conducted experiments based on existing publicly available datasets and resources. Then, we propose to regard LLMs as *Collaborators* to address the challenge of knowledge demands in CKRG, which may introduce biased or harmful information generated by LLMs. This issue is an inherent drawback of LLMs instead of coming from our approach. Additionally, all annotators involved in this work are paid. Finally, we disclose that the image icons in Figure 1 and Figure 2 are sourced from `https://icons8.com` under official free license.

References

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, Ben He, Shanshan Jiang, and Bin Dong. 2024. ChatGPT is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3098–3110, Torino, Italia. ELRA and ICCL.

Hyungjoo Chae, Yongho Song, Kai Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. 2023. Dialogue chain-of-thought distillation for commonsense-aware conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5606–5632, Singapore. Association for Computational Linguistics.

Liang Chen, Yang Deng, Yatao Bian, Zeyu Qin, Bingzhe Wu, Tat-Seng Chua, and Kam-Fai Wong. 2023a. Beyond factuality: A comprehensive evaluation of large language models as knowledge generators. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6325–6341, Singapore. Association for Computational Linguistics.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023b. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520, Singapore. Association for Computational Linguistics.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10602–10621, Singapore. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Amila Ferron, Amber Shore, Ekata Mitra, and Ameeta Agrawal. 2023. MEEP: Is this engaging? prompting large language models for dialogue evaluation in multilingual settings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2078–2100, Singapore. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Silin Gao, Jena D. Hwang, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. 2022. ComFact: A benchmark for linking contextual commonsense knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1656–1675, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2024. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18126–18134. AAAI Press.

Yutong Hu, Quzhe Huang, Mingxu Tao, Chen Zhang, and Yansong Feng. 2024. Can perplexity reflect large language model's ability in long text understanding? *arXiv preprint arXiv:2405.06105*.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. In *International Conference on Learning Representations*.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. SODA: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949, Singapore. Association for Computational Linguistics.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5203–5217, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yu Li, Baolin Peng, Yelong Shen, Yi Mao, Lars Liden, Zhou Yu, and Jianfeng Gao. 2022. Knowledge-grounded dialogue generation with a unified knowledge representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 206–218, Seattle, United States. Association for Computational Linguistics.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, and Chongyang Tao. 2024. Leveraging large language models for nlg evaluation: A survey. *arXiv preprint arXiv:2401.07103*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023a. Lost in the middle: How language models use long contexts. *CoRR*, abs/2307.03172.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing, pages 2511–2522, Singapore. Association for Computational Linguistics.

Zihan Liu, Mostofa Patwary, Ryan Prenger, Shrimai Prabhumoye, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Multi-stage prompting for knowledgeable dialogue generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1317–1337, Dublin, Ireland. Association for Computational Linguistics.

Hua Lu, Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2023. Towards boosting the open-domain chatbot with human feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4060–4078, Toronto, Canada. Association for Computational Linguistics.

Xuanfan Ni, Hongliang Dai, Zhaochun Ren, and Piji Li. 2023. Multi-source multi-type knowledge exploration and exploitation for dialogue generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12522–12537, Singapore. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

Revanth Reddy, Hao Bai, Wentao Yao, Sharath Chandra Etagi Suresh, Heng Ji, and ChengXiang Zhai. 2023. Social commonsense-guided search query generation for open-domain knowledge-powered conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 873–885, Singapore. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *Preprint*, arXiv:2307.11019.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Kurt Shuster, Mojtaba Komeili, Leonard Adolphs, Stephen Roller, Arthur Szlam, and Jason Weston. 2022. Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 373–393, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Yue Tan, Bo Wang, Anqi Liu, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2023. Guiding dialogue agents to complex semantic targets by dynamically completing knowledge graph. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6506–6518, Toronto, Canada. Association for Computational Linguistics.

Chen Tang, Hongbo Zhang, Tyler Loakman, Chenghua Lin, and Frank Guerin. 2023. Enhancing dialogue generation via dynamic graph knowledge aggregation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4604–4616, Toronto, Canada. Association for Computational Linguistics.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5624–5634, Florence, Italy. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ante Wang, Linfeng Song, Ge Xu, and Jinsong Su. 2023a. Domain adaptation for conversational query production with the RAG model feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9129–9141, Singapore. Association for Computational Linguistics.

Hongru Wang, Minda Hu, Yang Deng, Rui Wang, Fei Mi, Weichao Wang, Yasheng Wang, Wai-Chung Kwan, Irwin King, and Kam-Fai Wong. 2023b. Large language models as source planner for personalized knowledge-grounded dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9556–9569, Singapore. Association for Computational Linguistics.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2023c. Knowledge editing for large language models: A survey. *Preprint*, arXiv:2310.16218.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2020. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5811–5820, Online. Association for Computational Linguistics.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *CoRR*, abs/2402.19473.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peixiang Zhong, Yong Liu, Hao Wang, and Chunyan Miao. 2021. Keyword-guided neural conversational model. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14568–14576. AAAI Press.

Han Zhou, Xinchao Xu, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Siqi Bao, Fan Wang, and Haifeng Wang. 2022a. Link the world: Improving open-domain conversation with dynamic spatiotemporal-aware knowledge. *arXiv preprint arXiv:2206.14000*.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2021. Commonsense-focused dialogues for response generation: An empirical study. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 121–132, Singapore and Online. Association for Computational Linguistics.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2022b. Think before you speak: Explicitly generating implicit commonsense knowledge for response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1237–1252, Dublin, Ireland. Association for Computational Linguistics.

A Experimental Details

A.1 Implementation Details of Models

Here, we illustrate the implementation details of baselines and backbone models:

A.1.1 Baselines

This work studies LLMs-based CKRG, and the proposed approach is validated by instructing LLMs without supervised training. Thus, we consider the following baselines that can work in the same settings:

Vanilla It leverages the RG prompt reported in Table 12 to drive LLM to generate the response based on the dialogue history.

Cosmo (Kim et al., 2023) We adopt the 3B pretrained version of Cosmo. According to the input requirements, we set the temperature as 0.7 and use the following prompt for inference:

- *Situation:* Two participants are engaging in a friendly open-domain conversation, which can encompass a wide variety of topics without an explicit dialogue goal to be met within the conversation.
- *Instruction:* As a participant in the conversation, you should first try to understand the dialogue history and then generate the next dialogue response.

Since the Cosmo model only supports English, we only evaluate it on the English *DailyDialog* dataset.

MSDP (Liu et al., 2022) We adopt the officially proposed prompts to ask LLM to generate knowledge and responses with 10-shot settings. Since *MSDP* requires to dynamically construct the in-context learning examples for each dialogue context, we first conduct the *Hard Matching* (Zhou et al., 2022b) on the train and valid set of datasets to gain the commonsense knowledge aligned (H, K, R) triplets. Then, we use the Sentence-BERT [7] model (Reimers and Gurevych, 2019) as the embedding model to select the similar triplets as in-context learning examples based on the cosine similarity.

T-RAG (Wu et al., 2020; Gao et al., 2022) It employs all the mentioned entities as queries to retrieve commonsense knowledge from the ConceptNet (Speer et al., 2017). Then, the CKRG prompt reported in Table 13 will be used to instruct LLM to generate the response based on the dialogue history and the retrieved knowledge.

QP-RAG Compared to *T-RAG*, it first asks LLM to generate queries based on the dialogue history to retrieve knowledge. Following the recent query production works (Wang et al., 2023a; Reddy et al., 2023) that instruct LLM to generate queries as the competitive opponent, we implement this approach using our query production prompt reported in Table 14 to strictly compare the difference between this simple transfer idea and our method.

SCG-QP (Reddy et al., 2023) We employ the officially proposed approach to instruct LLM to generate queries for retrieving knowledge, and the other settings are the same as for *T-RAG*.

A.1.2 Backbone Models

We mainly consider the GPT-3.5 (`gpt-3.5-turbo-1106`) model as the backbone LLM to implement our *DCRAG* and baselines through iterative queries to OpenAI API interface[8]. To comprehensively investigate the LLM sensitivity of each method, we also employ the LLaMA3-8B (`Meta-Llama3-8B-Instruct`) and LLaMA3-70B (`Meta-Llama-3-70B-Instruct`) models to serve as the backbone LLM in our *LLM Sensitivity* experiment (see Sec 4.3). For inference, LLaMA3-8B and LLaMA3-70B use the API provided by GroqCloud[9]. All models are running with a 0.7 temperature setting.

A.2 The explanation of PPL

We report the *Perplexity (PPL)* (Jelinek et al., 1977) as a reference in Table 1 but do not compare this metric, since 1) *PPL* can measure the fluency of response, where lower is generally better. However, lower *PPL* is not equal to human-like (Kuribayashi et al., 2021). Thus, determining the appropriate range for *PPL* is challenging; 2) LLMs already excel in generating fluent text (Zhao et al., 2023), making it difficult to discern differences in fluency between responses generated by different methods using the same backbone LLM. As shown in Table 1, when using GPT-3.5 as the backbone model, the differences in *PPL* are significantly small. Previous work (Hu et al., 2024) has also highlighted the limitations of *PPL* in the context of LLMs.

[7] 970M, `paraphrase-multilingual-mpnet-base-v2`.

[8] `https://api.openai.com/v1/chat/completions`

[9] `https://groq.com/`

A.3 Details of Human Evaluation

We conduct a pair-wise human evaluation to ask three well-educated Chinese native-speaker annotators to compare the responses generated by *DCRAG* and baseline models. Mainly following the settings of (Zhong et al., 2022), we focus on four criteria:

- *Naturalness:* Which response sounds more natural and human-like?
- *Coherence:* Which response more logically follows the dialogue history?
- *Engagingness:* Which response can foster continued interaction and maintain or elevate interest in the conversation?
- *Informativeness:* Which response is more knowledgeable and contains more sufficient and rational information?

A.4 Details of RAGAs Evaluation

We use Retrieval Augmented Generation Assessment (RAGAs) (Es et al., 2024) to evaluate the RAG pipelines, including our *DCRAG* and the RAG-based CRKG baselines, based on the following four metrics:

- *Context Relevance:* Measure the relevance between the retrieved knowledge and the dialogue history.
- *Faithfulness:* Measure whether the generated response is faithful to the retrieved knowledge.
- *Answer Relevance:* Measure the relevance between the generated response and the dialogue history.
- *GeoMean:* The geomean of aforementioned metrics to measure the overall performance.

In practice, we configure the LLM evaluator and embedding model using `gpt-3.5-turbo-1106` and `text-embedding-3-small`, respectively, and follow the official settings in evaluation.

B The Impact of Facts Number

We also study the impact of the facts number on the quality of responses using our *DCRAG* and *T-RAG* on the *DailyDialog* dataset. The results are shown in the Figure 6. It can be seen that *DCRAG* and *T-RAG* show similar trends with the increased facts number, and the overall performance reaches the peak when using 20 facts to generate responses. Fewer facts are hard to augment the response generation, while more facts would bring more irrelevant information and challenge the long context capabilities of LLMs (Liu et al., 2023a). Thus, a moderate facts number is crucial.

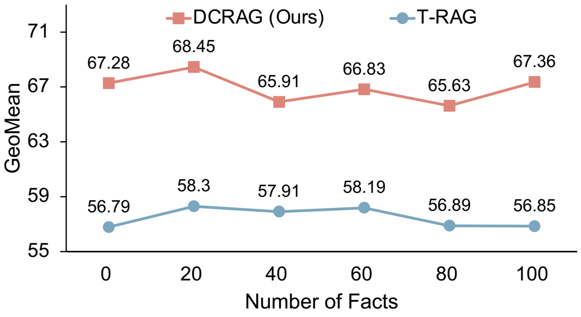


Figure 6: The impact of facts number.

C More Case Study

In Table 5 and Table 6, we report two cases from English *DailyDialog* dataset and Chinese *Diamante* dataset, respectively. The responses generated by *Vanilla* models tend to be uninformative and generic. In contrast, both *MSDP* and *T-RAG* generate more interesting and informative responses by leveraging externalized or retrieved commonsense knowledge. Differently, our *DCRAG* first performs deep reasoning to identify knowledge demands in a *Thought-then-Generate* way (e.g., Thoughts for Revised Explicit Queries and corresponding Queries). As a result, *DCRAG* achieves the best performance in generating high-quality responses.

D Error Analysis

In Table 7-Table 11, we report the real examples and corresponding analyses for each error type. The errors in *Query Production*, *Topic Planning*, and *Cross Revision* denote the generated queries or topics against their definitions; for example, one explicit query is not mentioned in the dialogue history. The errors in *Fact Retrieval* and *Fact Selection* denote the most facts that are irrelevant. The errors in *Response Generation* denote the response can not reply to the previous dialogue.

E Prompts

E.1 Response Generation

The prompts used in the response generation (RG) and commonsense knowledge-grounded response

Dialogue History *A:* I have to go up to London for a couple of days next week. Would you like to come?
Vanilla *R:* That sounds like fun! I'd love to join you.
MSDP *K:* (day, RelatedTo, nice) *R:* That sounds like a great opportunity! I'll have to check my schedule, but I'm definitely interested.
T-RAG *K:* (weeks, RelatedTo, week); (want to, RelatedTo, like to); (couple, HasContext, physics); (day, RelatedTo, language); (day, RelatedTo, cycle); (westminster, RelatedTo, london); (coupley, DerivedFrom, couple); ... *R:* That sounds like a great opportunity! Unfortunately, I won't be able to join you this time.
DCRAG *Thoughts for Revised Explicit Queries:* The explicit query entities are relevant to the discussion of the upcoming trip to London and the invitation to accompany the speaker. These need to be maintained to ensure coherence in the dialogue. *Revised Explicit Queries:* go up to London; couple of days; next week; Would you like to come *Thoughts for Revised Implicit Queries:* The implicit query entities are centered around the availability, interest, and schedule of the other person, which are essential for the dialogue to progress smoothly. *Revised Implicit Queries:* availability; interest; schedule *Thoughts for Revised Maintained Topics:* The maintained topic entities should continue to focus on the upcoming trip to London, the invitation, and the other person's availability and interest. These are important for maintaining coherence in the conversation. *Revised Maintained Topics:* go up to London; next week; availability; interest *Thoughts for Revised Extended Topics:* The extended topic entities can be expanded to include discussing potential itineraries or places to visit in London, transportation options, and accommodation arrangements, as these are relevant to the upcoming trip and can enhance the conversation. *Revised Extended Topics:* itineraries in London; places to visit; transportation options; accommodation arrangements *K:* (day, PartOf, week); (schedule, RelatedTo, activity); (person, CapableOf, visit place); (westminster, RelatedTo, london); (london, RelatedTo, capital city); (gallery, AtLocation, london); (oxford street, AtLocation, london); ... *R:* That sounds great! I'd love to come with you. Do you already have an itinerary in mind for the trip?

Table 5: Case Study #1 on English *DailyDialog* dataset. K and R denote externalized/retrieved knowledge and response. We also report the outputs of entities and their corresponding thoughts in the *Cross Revision* stage of *DCRAG*. In this context, green represents the queries, and purple denotes the topics.

generation (CKRG) models are reported in Table 12 and Table 13 respectively.

E.2 DCRAG

The prompts of our proposed *DCRAG* are reported in Table 14 (*Query Production*), Table 15 (*Topic Planning*), Table 16 (*Cross Revision*), Table 17 (*Fact Selection*), and Table 18 (*Response Generation*) respectively.

E.3 GPT-4 Evaluation

The prompts used in GPT-4 Evaluation are reported in Table 19 and Table 20.

Dialogue History

A: 好厉害！我也就煮个面条啥的，我是不是买点面条就行？ (So impressive! I just cook noodles or something. Should I just buy some noodles?)

B: 买挂面，买个不粘锅，就可以了。别忘了买些蔬菜，只吃面条可不行。 (Buy some dried noodles and a non-stick pan, and you're good to go. Don't forget to buy some vegetables; just eating noodles isn't enough.)

A: 蔬菜要买什么？怎么做？ (What vegetables should I buy? How do I prepare them?)

B: 你喜欢吃什么了！吃面条多的话就买点西红柿、鸡蛋、蒜苔。 (It depends on what you like! If you eat a lot of noodles, get some tomatoes, eggs, and garlic shoots.)

A: 我想吃带馅儿的！ (I want to eat something with filling!)

Vanilla

R: 那你可以试试做饺子,自己包馅儿,很好吃的 (Then you can try making dumplings and wrapping the filling yourself; they taste great.)

MSDP

K: (带馅儿 Stuffed, IsA, 饺子 Dumplings); (带馅儿 Stuffed, IsA, 馄饨 Wontons); (带馅儿 Stuffed, IsA, 饺子皮 Dumpling wrappers); (饺子皮 Dumpling wrappers, MadeOf, 面粉 Flour); (蒜苔 Garlic shoots, IsA, 蔬菜 Vegetables)

R: 那就买点饺子皮和馅料，自己包一些饺子吧，很有趣的厨艺体验。 (Then buy some dumpling wrappers and filling, and make some dumplings yourself. It's a fun cooking experience.)

T-RAG

K: (面条 Noodles, UsedFor, 自己煮面 Cooking noodles by oneself); (吃面 Eating noodles, Causes, 配小菜 Accompanied by side dishes); (吃面条 Eating noodles, Causes, 饱 Fullness); (去买面 Buying noodles, MotivatedByGoal, 吃面 Eating noodles); (吃面条 Eating noodles, MotivatedByGoal, 肚子饿 Hunger); ...

R: 那你可以买点猪肉馅儿,我们一起包饺子吧！ (Then you can buy some minced pork, and we can make dumplings together!)

DCRAG

Thoughts for Revised Explicit Queries: The main focus is on buying ingredients for cooking noodles, so we should maintain the explicit query entities related to noodles and cooking, but we can remove the explicit query entities related to other dishes.

Revised Explicit Queries: 面条; 挂面; 不粘锅 (Noodles; Dried noodles; Non-stick pan)

Thoughts for Revised Implicit Queries: Since the main topic is on making dumplings, we can consider adding implicit query entities related to different types of dumpling fillings and dumpling making methods.

Revised Implicit Queries: 不同种类的饺子馅料; 饺子的制作方法 (Different types of dumpling fillings; Dumpling making methods)

Thoughts for Revised Maintained Topics: The main topic of making dumplings is still relevant, so we should maintain the topic entities related to noodles, vegetables, and dumplings.

Revised Maintained Topics: 面条; 蔬菜; 西红柿; 鸡蛋; 蒜苔 (Noodles; Vegetables; Tomatoes; Eggs Garlic shoots)

Thoughts for Revised Extended Topics: Since the interest is in making dumplings, we should extend the topic to include different types of dumpling fillings, cooking methods for making dumplings, and possibly dipping sauces for dumplings.

Revised Extended Topics: 不同种类的饺子馅料; 饺子的制作方法; 饺子蘸料的制作方法 (Different types of dumpling fillings; Dumpling making methods; Dipping sauce methods for dumplings)

K: (馅料 Filling, AtLocation, 饺子 Dumplings); (韭菜 Chives, AtLocation, 饺子 Dumplings); (面条 Noodles, UsedFor, 自己煮面 Cooking noodles by oneself); (鸡丝面 Shredded chicken noodles, IsA, 面条 Noodles); (面条 Noodles, PartOf, 炒面 Fried noodles); ...

R: 你可以尝试做饺子馅儿,比如韭菜猪肉馅儿或者虾仁馅儿,再配上一些醋和姜丝,味道会很不错的。 (You can try making dumpling fillings, such as chive and pork or shrimp, and pair them with some vinegar and shredded ginger for a great taste.)

Table 6: Case Study #2 on Chinese *Diamante* dataset. K and R denote externalized/retrieved knowledge and response. We also report the outputs of entities and their corresponding thoughts in the *Cross Revision* stage of *DCRAG*.

Explicit Queries Error

Dialogue History
A: What's wrong with you? You look pale.
B: I'm not sure, I feel hot and cold.
A: When did the trouble start?

Thoughts for Explicit Queries: The explicit query entities revolve around the speaker's health condition and symptoms.
Explicit Queries: health condition; symptoms; trouble

Analysis: The explicit queries of 'health condition' and 'symptoms' are did not explicitly mentioned in dialogue history, which conflicts with the definition of explicit queries.

Implicit Queries Error

Dialogue History
A: What's she doing?
B: She's sitting under the tree.
A: Is Tim in the garden, too?
B: Yes, he is. He's climbing the tree.
A: I beg your pardon? Who's climbing the tree.

Thoughts for Implicit Queries: The implicit query entities are the location of the garden, the activity of climbing, and the concept of who is involved in the dialogue.
Implicit Queries: garden; climbing; who

Analysis: The implicit queries, including 'garden', 'climbing', and 'who', are all explicitly mentioned in dialogue history, which conflicts with the definition of implicit queries.

Table 7: Examples of errors in Query Production. Red represents the error entity, while blue represents the corresponding analysis.

Maintained Topics Error

Dialogue History
A: How do you research your books?
B: I always visit the places that I use as setting for stories. Readers like things to be as factually correct as possible-even in fiction! I usually base my characters on people I have met. Most character are a mixture of the characters of two or more people.
A: I really like your books. I've just finished this one. I read it from cover to cover in a single day. Congratulation on making the bestseller list yet again!
B: Thank you. Let met sign it for you... there you go.
A: Oh! Thank you very much. I heard that you are currently writing a collection of short stories. What are the stories about?

Thoughts for Maintained Topics: The topic of writing process, inspiration for characters, and reader preferences can be maintained as they are currently being discussed and are of interest to the participants.
Maintained Topics: writing process; inspiration for characters; reader preferences

Analysis: A wonder what the content of the collection of short stories is about, while the 'writing process', 'inspiration for characters', and 'reader preferences' have already been discussed in the previous conversation, so it does not need to be involved in the next response.

Extended Topics Error

Dialogue History
A: Really? Is he married?
B: No, of course not. He is still single.
A: Then what's your problem?
B: I try to keep it to myself. But there is a lot of gossip about us.
A: Oh, I see. Office romance tends to be the subject of gossip.

Thoughts for Extended Topics: An extended topic entity could be 'professionalism in the workplace', as it is related to the discussion of office romance and handling workplace gossip. Additionally, 'impact of gossip on work relationships' can be an extended topic entity, providing a broader perspective on the issue.
Extended Topics: professionalism in the workplace; impact of gossip on work relationships

Analysis: We can see that this dialogue history mainly talks about the office gossip. Thus, 'professionalism in the workplace' is irrelevant to the current dialogue context and does not need to be extended in the next response.

Table 8: Examples of errors in Topic Planning. Red represents the error entity, while blue represents the corresponding analysis.

Revised Explicit Queries Error
Dialogue History *A:* It’s quiet everywhere in winter.
Thoughts for Explicit Queries: The explicit query entities include ’quiet’ and ’winter’ as they are meaningful words mentioned in the dialogue history. *Explicit Queries:* quiet; winter
Thoughts for Revised Explicit Queries: The revised explicit query entities should focus on the feeling of quietness and its impact during winter, such as ’winter activities’ and ’emotions in winter’. *Revised Explicit Queries:* quietness; winter activities; emotions in winter
Analysis: All entities in *Revised Explicit Queries* are not explicitly mentioned in dialogue history. We notice that the entities in *Explicit Queries* are correct, suggesting the backbone has made an excessive inference.
Revised Implicit Queries Error
Dialogue History *A:* What’s she doing? *B:* She’s sitting under the tree. *A:* Is Tim in the garden, too? *B:* Yes, he is. He’s climbing the tree. *A:* I beg your pardon? Who’s climbing the tree.
Thoughts for Implicit Queries: The implicit query entities are the location of the garden, the activity of climbing, and the concept of who is involved in the dialogue. *Implicit Queries:* garden; climbing; who
Thoughts for Revised Implicit Queries: The revised implicit query entities should include the concept of ’pardon’ and the location of the tree for climbing. *Revised Implicit Queries:* pardon; tree
Analysis: All entities in *Revised Implicit Queries* and *Implicit Queries* are incorrect since they are explicitly mentioned in the dialogue history. It shows a case of error propagation.
Revised Maintained Topics Error
Dialogue History *A:* Tom, I won first prize in the poetry contest.
Thoughts for Maintained Topics: The maintained topic entities are ’Tom’ and ’poetry contest’ because they are the main focus of the conversation, and they provide context for the achievement of winning the first prize. *Maintained Topics:* Tom; poetry contest
Thoughts for Revised Maintained Topics: The maintained topic entities are ’Tom’ and ’poetry contest’ because they are the main focus of the conversation, and they provide context for the achievement of winning the first prize. *Revised Maintained Topics:* Tom; poetry contest; impact of the win; upcoming opportunities
Analysis: The topics of ’impact of the win’ and ’upcoming opportunities’ are more suitable as extended topics rather than maintained topics because it is difficult to infer from the dialogue history.
Revised Extended Topics Error
Dialogue History *A:* How are you today?
Thoughts for Extended Topics: Since the dialogue history is limited to a simple greeting, we can extend the topic with related questions such as ’How was your day?’ or ’Did anything interesting happen?’ to keep the conversation engaging and interactive. *Extended Topics:* How was your day?; Did anything interesting happen?
Thoughts for Revised Extended Topics: The extended topic entities ’How was your day?’ and ’Did anything interesting happen?’ are suitable continuations of the greeting and can maintain an engaging conversation. No revisions are needed. *Revised Extended Topics:* How was your day?; Did anything interesting happen?
Analysis: The topic of ’How was your day?’ is only the rewriting of dialogue history rather than an inferred topic. We can see that this error occurs in the earlier *Extended Topics*.

Table 9: Examples of errors in Cross Revision. Red represents the error entity, while blue represents the corresponding analysis.

Foreseen Set Error
Dialogue History *A:* The hotel operator. Anything I can do for you? *B:* I've been trying to get through to the booking office for the last ten minutes. But the line has been engaged all the time. Why is it taking so long? *A:* I'm sorry to hear that. I'm afraid it is often the case during the peak hours. Shall I ask them to call you when the line is free? *B:* OK. Please do so if it is not troublesome. *A:* No trouble at all. May I know your name and room number?
Foreseen Set: (solution, RelatedTo, service); (solution, IsA, method); (office, RelatedTo, service); (customer service, RelatedTo, customer); (requests, RelatedTo, request); (hours, RelatedTo, hour); (customer, RelatedTo, service); (hours, FormOf, hour); (office, Synonym, service); (peak hours, FormOf, peak hour); ...
Analysis: Most facts in *Foreseen Set* is irrelevant to current dialogue context.
Unforeseen Set Error
Dialogue History *A:* Well, I'm afraid I must be going now. I have no idea it was so late.
Unforeseen Set: (hydrolaparoscopy, RelatedTo, solution); (cleave, RelatedTo, impact); (late bloomer, DerivedFrom, late); (time, RelatedTo, hour minute); (staying in bed, Causes, late); (desolvation, RelatedTo, solution); (time of departure, RelatedTo, time); (plan, Antonym, wing it); (tipping point, RelatedTo, time); (killing spree, RelatedTo, time); ...
Analysis: similarly, most facts in *Unforeseen Set* is irrelevant and not useful. It is the most common error.
Selected Facts Error
Dialogue History *A:* How are you? *B:* I'm doing well. *A:* How may I help you?
Selected Facts: (support, RelatedTo, assistance); (health, IsA, wellbeing); (support, RelatedTo, encouragement); (support, RelatedTo, peer); (guide dog, RelatedTo, assistance); (taking walk, UsedFor, health); (health, HasProperty, very important to maintaining happiness); (autopilot, RelatedTo, assistance); (relieve, RelatedTo, assistance); ...
Analysis: Despite LLM's considerable effort to select useful facts, *Selected Facts* in this case is still irrelevant to the current dialogue context.

Table 10: Examples of errors in Fact Retrieval and Fact Selection. Red represents the error facts, while blue represents the corresponding analysis.

Generated Response Error
Example #1 (Out of character)
Dialogue History *A:* Nani Book Store. How can I help you?
Response: What kind of books are you interested in? We have a wide range of genres available at Nani Book Store.
Analysis: This response clearly mistook its own character; it merely continued the dialogue history instead of answering the questions within it. This is due to the LLM's misunderstanding.
Example #2 (Contrary to commonsense)
Dialogue History *A:* Could you please wait for a minute? I'll buy the tickets *B:* All right. *A:* Here is your ticket.
Response: Thank you! I'm excited about the ferry ride. Do you know if there are any discounts or promotions for the tickets?
Analysis: The error in this response is more complex because the tickets have already been obtained; asking again about the discount on tickets would violate the sequence of events and be contrary to commonsense.

Table 11: Examples of errors in Response Generation. Red represents the error response, while blue represents the corresponding analysis.

```
[Task Description]
You will be participating in an open-domain dialogue generation (Chit-Chat) task, which can encompass
a wide variety of topics without an explicit dialogue goal to be met within the conversation. This
conversation is two-party and in {Language}. You should act as my friend (denoted as 'Friend') and
engage with me (denoted as 'Me'). You should first try to understand the current dialogue context by
reading the 'Dialogue History' in the [Input] and then continue the dialogue by generating the next
'Dialogue Response' in the [Output] from my friend's perspective. You should follow the [Guidance]
to generate a dialogue response. Your output format should follow the [Output Format].

[Guidance]
You should strictly follow my guidance:
1. To continue the conversation, you must produce a natural, coherent, engaging, and informative
response based on the given dialogue history.
2. Your response must reply to the last dialogue utterance of "{Last Turn}".
3. Your response must be 1-3 sentences and in {Language}.
4. You should strictly follow the given output format and can't output other information.
If you break my guidance, you will be penalized.

[Output Format]
Your output should strictly follow a JSON format and can be directly decoded by Python. Here is an
example:
{"Dialogue Response": "Friend-{Target Turn Index}: {Response Example}"}

[Input]
{Input}

[Output]
```

Table 12: Prompt of RG models in our experiments. The $\{\cdot\}$ represents the metadata for specific case, while $\{\cdot\}$ represents the input content.

```
[Task Description]
You will be participating in an open-domain dialogue generation (Chit-Chat) task, which can encompass
a wide variety of topics without an explicit dialogue goal to be met within the conversation. This
conversation is two-party and in {Language}. You should act as my friend (denoted as ’Friend’) and
engage with me (denoted as ’Me’). You should first try to understand the current dialogue context by
reading the ’Dialogue History’ in the [Input] and then continue the dialogue by generating the next
’Dialogue Response’ in the [Output] from my friend’s perspective. You should follow the [Guidance]
and [Criteria] to generate a dialogue response. Your output format should follow the [Output Format].

[Guidance]
You should strictly follow my guidance:
1. Contextual knowledge is a set of commonsense knowledge that is explained in the [Criteria].
2. Contextual knowledge is retrieved from the external knowledge base via a retrieval system.
However, irrelevant or noisy knowledge may be included. Therefore, before using knowledge to
generate a response, it’s crucial to assess the relevance and usefulness of each piece of contextual
knowledge and select the most useful knowledge. If you believe there is no useful knowledge available,
you have the choice to ignore the contextual knowledge to generate a response directly or use your
personal commonsense, whichever is better.
3. To continue the conversation, you must produce a natural, coherent, engaging, and informative
response based on the given dialogue history and selected useful commonsense knowledge.
4. Your response must reply to the last dialogue utterance of "{Last Turn}".
5. Your response must be 1-3 sentences and in {Language}.
6. You should strictly follow the given output format and can’t output other information.
If you break my guidance, you will be penalized.

[Criteria]
Commonsense Knowledge: It consists of facts about the everyday world, such as "Lemons are sour",
that all humans are expected to know. In this task, each of commonsense knowledge is structured
as "(Head Entity, Relation, Tail Entity)". For example, the fact (Sunflower, IsA, Flower) conveys
that a sunflower is a kind of flower. Useful commonsense knowledge is important since it provides
additional information for corresponding dialogue that can be used to generate a more interesting
and informative response.

[Output Format]
Your output should strictly follow a JSON format and can be directly decoded by Python. Here is an
example:
{"Dialogue Response": "Friend-{Target Turn Index}: {Response Example}"

[Input]
{Input}}

[Output]
```

Table 13: Prompt of CKRG models in our experiments. The {·} represents the metadata for specific case, while {·} represents the input content.

```
[Task Description]
You will be participating in an open-domain dialogue generation (Chit-Chat) task, which can encompass
a wide variety of topics without an explicit dialogue goal to be met within the conversation. This
conversation is two-party and in {Language}. Your task is to generate the query entities for the
given dialogue history. Specifically, after understanding the current dialogue history by reading
the "Dialogue History" in the [Input], create brief descriptions for the thoughts of explicit and
implicit query entities in the "Thoughts for Explicit Query Entities" and "Thoughts for Implicit
Query Entities" in the [Output]. Then, generate the "Explicit Query Entities" and "Implicit Query
Entities" in the [Output] from the perspective of dialogue history. You should follow the [Guidance]
and [Criteria] to generate thoughts for explicit query entities, explicit query entities, thoughts
for implicit query entities, and implicit query entities.  Your output format should follow the
[Output Format].

[Guidance]
You should strictly follow my guidance:
1. You should read the [Criteria] to understand the Explicit Query Entities, Implicit Query Entities,
The Use of Query Entities, and Commonsense Knowledge.
2. You should think step by step about which entities are explicit or implicit query entities and
why these entities can be explicit or implicit query entities. You should give several sentences to
describe your thoughts. Then, you should extract the {Language} explicit query entities and extract
the {Language} implicit query entities from the perspective of dialogue history.
3. You should strictly follow the given output format and can't output other information.
If you break my guidance, you will be penalized.

[Criteria]
1.  Explicit Query Entities: These can be meaningful words or phrases that are mentioned in the
dialogue history.  To extract the explicit query entities, you may need to check each dialogue
utterance and consider what has been discussed and mentioned.
2.  Implicit Query Entities: These can be words, phrases, or sentences that are not explicitly
mentioned in the dialogue history.  To extract the implicit query entities, you may need certain
reasoning based on the dialogue history and explicit query entities, including understanding the
dialogue utterances to summarize the key information, extending the involved topics that have been
mentioned, identifying the intentions or interests of each dialogue participant, and so on.
3. The Use of Query Entities: The query entities will serve as starting nodes to retrieve the related
commonsense knowledge from the external knowledge graph.
4. Commonsense Knowledge: It consists of facts about the everyday world, such as "Lemons are sour",
that all humans are expected to know.  In this task, each of commonsense knowledge is structured
as "(Head Entity, Relation, Tail Entity)". For example, the fact (Sunflower, IsA, Flower) conveys
that a sunflower is a kind of flower. Useful commonsense knowledge is important since it provides
additional information for corresponding dialogue that can be used to generate a more interesting
and informative response.

[Output Format]
Your output should strictly follow a JSON format and can be directly decoded by Python. Here is an
example:
{"Thoughts for Explicit Query Entities": "Your brief thoughts for explicit query entities", "Explicit
Query Entities": [Explicit query entities], "Thoughts for Implicit Query Entities": "Your brief
thoughts for implicit query entities", "Implicit Query Entities": [Implicit query entities]}

[Input]
{Input}


[Output]
```

Table 14: Prompt of *Query Production* in *DCRAG*. The {·} represents the metadata for specific case, while {·} represents the input content.

[Task Description]
You will be participating in an open-domain dialogue generation (Chit-Chat) task, which can encompass a wide variety of topics without an explicit dialogue goal to be met within the conversation. This conversation is two-party and in {Language}. Your task is to generate the topic entities for the next dialogue response. Specifically, after understanding the current dialogue history by reading the "Dialogue History" in the [Input], create brief descriptions for the thoughts of maintained and extended topic entities in the "Thoughts for Maintained Topic Entities" and "Thoughts for Extended Topic Entities" in the [Output]. Then, generate the "Maintained Topic Entities" and "Extended Topic Entities" in the [Output] from the perspective of the next dialogue response. You should follow the [Guidance] and [Criteria] to generate thoughts for maintained topic entities, maintained topic entities, thoughts for extended topic entities, and extended topic entities. Your output format should follow the [Output Format].

[Guidance]
You should strictly follow my guidance:
1. You should read the [Criteria] to understand the Maintained Topic Entities, Extended Topic Entities, The Use of Topic Entities, and Commonsense Knowledge.
2. You should think step by step about which entities are maintained or extended topic entities and why these entities can be maintained or extended topic entities. You should give several sentences to describe your thoughts. Then, you should plan the {Language} maintained topic entities and the {Language} extended topic entities from the perspective of the next dialogue response.
3. You should strictly follow the given output format and can't output other information.
If you break my guidance, you will be penalized.

[Criteria]
1. Maintained Topic Entities: These can be specific words or phrases that are the main topic currently being discussed and should be a part of the next dialogue response, considering the coherent and engaging interaction. To plan the maintained topic entities, you need to understand each dialogue utterance to identify the topic shift flow in this dialogue.
2. Extended Topic Entities: These can be specific words, phrases, or sentences that are not mentioned in the dialogue history, but the next diverse and informative dialogue response may be involved. To plan the extended topic entities, you need to confirm the existing topics in this dialogue, then conduct commonsense reasoning to infer the new topics, such as "For existing A, it will affect B and C can solve A. Thus, B and C can be as the extended topics".
3. The Use of Topic Entities: The topic entities are crucial to dialogue response planning. On the one hand, these entities (especially for the maintained topic entities) guide a coherent and engaging dialogue response. On the other hand, they (especially for the extended topic entities) can also serve as the ending nodes, directing the commonsense knowledge and controlling the scope of retrieved knowledge.
4. Commonsense Knowledge: It consists of facts about the everyday world, such as "Lemons are sour", that all humans are expected to know. In this task, each of commonsense knowledge is structured as "(Head Entity, Relation, Tail Entity)". For example, the fact (Sunflower, IsA, Flower) conveys that a sunflower is a kind of flower. Useful commonsense knowledge is important since it provides additional information for corresponding dialogue that can be used to generate a more interesting and informative response.

[Output Format]
Your output should strictly follow a JSON format and can be directly decoded by Python. Here is an example:
{"Thoughts for Maintained Topic Entities": "Your brief thoughts for maintained topic entities", "Maintained Topic Entities": [Maintained topic entities], "Thoughts for Extended Topic Entities": "Your brief thoughts for extended topic entities", "Extended Topic Entities": [Extended topic entities]}

[Input]
{Input}

[Output]

Table 15: Prompt of *Topic Planning* in *DCRAG*. The {·} represents the metadata for specific case, while {·} represents the input content.

[Task Description]
You will be participating in an open-domain dialogue generation (Chit-Chat) task, which can encompass a wide variety of topics without an explicit dialogue goal to be met within the conversation. This conversation is two-party and in {Language}. Your task is to generate the revised query entities and topic entities. Specifically, after understanding the current dialogue history by reading the "Dialogue History" in the [Input], create brief descriptions for the thoughts of revised query entities and topic entities in the "Thoughts for Revised Explicit Query Entities", "Thoughts for Revised Implicit Query Entities", "Thoughts for Revised Maintained Topic Entities", and "Thoughts for Revised Extended Topic Entities" in the [Output]. Then, generate the "Revised Explicit Query Entities", "Revised Implicit Query Entities", "Revised Maintained Topic Entities", and "Revised Extended Topic Entities" in the [Output] considering the relationship of query entities and topic entities. Your output format should follow the [Output Format].

[Guidance]
You should strictly follow my guidance:
1. You should read the [Criteria] to understand the Explicit Query Entities, Implicit Query Entities, The Use of Query Entities, Maintained Topic Entities, Extended Topic Entities, The Use of Topic Entities, and Commonsense Knowledge.
2. You should think step by step about which (query or topic) entities need to be added, retained, revised, or removed from the perspective of the relationship between the query entities and topic entities. The former ought to be in the dialogue history, and the latter ought to be in the response. You should give several sentences to describe your thoughts and generate them.
3. You should strictly follow the given output format and can't output other information.
If you break my guidance, you will be penalized.

[Criteria]
1. Explicit Query Entities: These can be meaningful words or phrases that are mentioned in the dialogue history. The explicit query entities are extracted by considering what has been discussed and mentioned.
2. Implicit Query Entities: These can be words, phrases, or sentences that are not explicitly mentioned in the dialogue history. The implicit query entities are extracted by conducting certain reasoning based on the dialogue history and explicit query entities, including understanding the dialogue utterances to summarize the key information, extending the topics involved in the entities, identifying the intentions or interests of each dialogue participant, and so on.
3. Maintained Topic Entities: These can be specific words or phrases that are the main topic currently being discussed and should be a part of the next dialogue response, considering the coherent and engaging interaction. The maintained topic entities are obtained by understanding each dialogue utterance to identify the topic shift flow in this dialogue.
4. Extended Topic Entities: These can be specific words, phrases, or sentences that are not mentioned in the dialogue history, but the next diverse and informative dialogue response may be involved. The extended topic entities are obtained by conducting commonsense reasoning to infer the new topics, such as "For existing A, it will affect B and C can solve A. Thus, B and C can be as the extended topics".
5. Commonsense Knowledge: It consists of facts about the everyday world, such as "Lemons are sour", that all humans are expected to know. In this task, each of commonsense knowledge is structured as "(Head Entity, Relation, Tail Entity)". For example, the fact (Sunflower, IsA, Flower) conveys that a sunflower is a kind of flower. Useful commonsense knowledge is important since it provides additional information for corresponding dialogue that can be used to generate a more interesting and informative response.

[Output Format]
Your output should strictly follow a JSON format and can be directly decoded by Python. Here is an example:
{"Thoughts for Revised Explicit Query Entities": "Your brief thoughts for revised explicit query entities", "Revised Explicit Query Entities": [Revised explicit query entities], "Thoughts for Revised Implicit Query Entities": "Your brief thoughts for revised implicit query entities", "Revised Implicit Query Entities": [Revised implicit query entities], "Thoughts for Revised Maintained Topic Entities": "Your brief thoughts for revised maintained topic entities", "Revised Maintained Topic Entities": [Revised maintained topic entities], "Thoughts for Revised Extended Topic Entities": "Your brief thoughts for revised extended topic entities", "Revised Extended Topic Entities": [Revised extended topic entities]}

[Input]
{Input}

[Output]

Table 16: Prompt of *Cross Revision* in *DCRAG*. The {·} represents the metadata for specific case, while {·} represents the input content.

```
[Task Description]
You will be participating in an open-domain dialogue generation (Chit-Chat) task, which can encompass
a wide variety of topics without an explicit dialogue goal to be met within the conversation. This
conversation is two-party and in {Language}. Your task is to select the top {Knowledge Number}
most useful knowledge from the retrieved knowledge. Specifically, after understanding the current
dialogue history by reading the "Dialogue History" in the [Input], you need to assess the relevance
and usefulness of each knowledge in the "Retrieved Knowledge" in the [Input]. Then, select the
top {Knowledge Number} most useful knowledge according to your assessment and output the result as
"Selected Knowledge" in the [Output]. You should follow the [Guidance] and [Criteria] to output the
selected knowledge. Your output format should follow the [Output Format].

[Guidance]
You should strictly follow my guidance:
1. Retrieved knowledge is a set of commonsense knowledge that is explained in the [Criteria].
2. You should output the selected top {Knowledge Number} knowledge based on your comprehensive
assessment. If the amount of knowledge in retrieved knowledge is less than {Knowledge Number}, you
just need to rank the knowledge in retrieved knowledge and output all of them.
3. You should strictly follow the given output format and can't output other information.
If you break my guidance, you will be penalized.

[Criteria]
1. Commonsense Knowledge: It consists of facts about the everyday world, such as "Lemons are sour",
that all humans are expected to know. In this task, each of commonsense knowledge is structured
as "(Head Entity, Relation, Tail Entity)". For example, the fact (Sunflower, IsA, Flower) conveys
that a sunflower is a kind of flower. Useful commonsense knowledge is important since it provides
additional information for corresponding dialogue that can be used to generate a more interesting
and informative response.
2. Retrieved Knowledge: It is retrieved from the external knowledge base via a retrieval system.
However, irrelevant or noisy knowledge may be included. Therefore, it's crucial to assess each
piece of retrieved knowledge, particularly in terms of its relevance to the dialogue history and
its potential usefulness for generating the next engaging and informative responses."

[Output Format]
Your output should strictly follow a JSON format and can be directly decoded by Python. Here is an
example:
{"Selected Knowledge": ["[1]-(head 1, relation 1, tail 1)", "[2]-(head 2, relation 2, tail 2)",
...]}

[Input]
{Input}


[Output]
```

Table 17: Prompt of *Fact Selection* in *DCRAG*. The {·} represents the metadata for specific case, while {·} represents the input content.

```
[Task Description]
You will be participating in an open-domain dialogue generation (Chit-Chat) task, which can encompass
a wide variety of topics without an explicit dialogue goal to be met within the conversation. This
conversation is two-party and in {Language}. You should act as my friend (denoted as ’Friend’) and
engage with me (denoted as ’Me’). You should first try to understand the current dialogue context by
reading the ’Dialogue History’ in the [Input] and then continue the dialogue by generating the next
’Dialogue Response’ in the [Output] from my friend’s perspective. You should follow the [Guidance]
and [Criteria] to generate a dialogue response. Your output format should follow the [Output Format].

[Guidance]
You should strictly follow my guidance:
1. Contextual knowledge is a set of commonsense knowledge that is explained in the [Criteria].
2.  Contextual knowledge is retrieved from the external knowledge base via a retrieval system.
However, irrelevant or noisy knowledge may be included.  Therefore, before using knowledge to
generate a response, it’s crucial to assess the relevance and usefulness of each piece of contextual
knowledge and select the most useful knowledge. If you believe there is no useful knowledge available,
you have the choice to ignore the contextual knowledge to generate a response directly or use your
personal commonsense, whichever is better.
3. To continue the conversation, you must produce a natural, coherent, engaging, and informative
response based on the given dialogue history and selected useful commonsense knowledge.
4. You can use the information of Explicit Query Entities, Implicit Query Entities, Maintained Topic
Entities, and Extended Topic Entities to help generate the response.
5. Your response must reply to the last dialogue utterance of "{Last Turn}".
6. Your response must be 1-3 sentences and in {Language}.
7. You should strictly follow the given output format and can’t output other information.
If you break my guidance, you will be penalized.

[Criteria]
1.  Explicit Query Entities: These can be meaningful words or phrases that are mentioned in the
dialogue history. The explicit query entities are extracted by considering what has been discussed
and mentioned.
2.  Implicit Query Entities: These can be words, phrases, or sentences that are not explicitly
mentioned in the dialogue history. The implicit query entities are extracted by conducting certain
reasoning based on the dialogue history and explicit query entities, including understanding the
dialogue utterances to summarize the key information, extending the topics involved in the entities,
identifying the intentions or interests of each dialogue participant, and so on.
3. Maintained Topic Entities: These can be specific words or phrases that are the main topic currently
being discussed and should be a part of the next dialogue response, considering the coherent and
engaging interaction.  The maintained topic entities are obtained by understanding each dialogue
utterance to identify the topic shift flow in this dialogue.
4. Extended Topic Entities: These can be specific words, phrases, or sentences that are not mentioned
in the dialogue history, but the next diverse and informative dialogue response may be involved. The
extended topic entities are obtained by conducting commonsense reasoning to infer the new topics,
such as "For existing A, it will affect B and C can solve A. Thus, B and C can be as the extended
topics".
5. Commonsense Knowledge: It consists of facts about the everyday world, such as "Lemons are sour",
that all humans are expected to know.  In this task, each of commonsense knowledge is structured
as "(Head Entity, Relation, Tail Entity)". For example, the fact (Sunflower, IsA, Flower) conveys
that a sunflower is a kind of flower. Useful commonsense knowledge is important since it provides
additional information for corresponding dialogue that can be used to generate a more interesting
and informative response.

[Output Format]
Your output should strictly follow a JSON format and can be directly decoded by Python. Here is an
example:
{"Dialogue Response": "Friend-{Target Index}: {Response Example}"}

[Input]
{Input}


[Output]
```

Table 18: Prompt of *Response Generation* in *DCRAG*. The {·} represents the metadata for specific case, while {·} represents the input content. Compared to the CKRG prompt in Table 13, it additionally adds thoughts and entities outputted by previous stages.

```
Score the following response given the corresponding dialogue context on a continuous scale from 0
to 100, where a score of zero means ‘disengaging’ and a score of 100 means ‘very engaging’. Assume
the response immediately follows the dialogue context. Consider that engagingness of a response is
defined by the following qualities: variety of response according to the context (such as responding
to ‘Hi how are you?’ with ‘I feel magnificent, because I just successfully defended my PhD! How are
you?’ instead of ‘Good, how are you?’), likelihood of encouraging the other participant to respond
(such as ‘I love legos! I like using them to make funny things. Do you like legos?’ instead of ‘I like
legos.’), likelihood of encouraging a quality response from the other participant, interestingness,
specificity, and likelihood of creating a sense of belonging for the other participant.
Dialogue context: {dialogue}
Response: {response}
Score:
```

Table 19: Prompt of the MEEP (Ferron et al., 2023) metric in our experiments. We employ the type of MEEP+SA here since it shows higher correlations with humans. {dialogue} and {response} is the corresponding dialogue history and response that needs to be evaluated.

```
[Task Description]
Here is a point-wise Dialogue Response Evaluation task. All [Input] are in {Language}. You are
required to act as a professional native-speaker human annotator to judge the given Dialogue Response
in [Input]. Your evaluation should follow the [Guidance] and [Criteria]. The output format should
follow the [Output Format].

[Guidance]
You should strictly follow my guidance:
1. You should first read the dialogue history and the response carefully.
2. You should rate the response for each aspect independently, according to the corresponding
criteria. A low score in one aspect should not influence another aspect.
3. Each score is between 1 (lowest) and 5 (highest) and should be an int score.
4. You should strictly follow the given output format and can’t output other information.
If you break my guidance, you will be penalized.

[Criteria]
1. Informativeness: Is the "Dialogue Response” knowledgeable and contains sufficient and rational
information? A high score for informativeness should indicate that the response offers novel,
detailed, accurate, and appropriate information that aligns with the participant’s needs.
2. Overall Score: How is the overall quality of the "Dialogue Response"? This score comprehensively
assesses whether the response can achieve a satisfying interaction.

[Output Format]
Your output should strictly follow the JSON format and can be directly decoded by Python. Here is
an example:
{”Informativeness”: [Your Score], ”Overall Score”: [Your Score]}

[Input]
{Input}

[Output]
```

Table 20: Prompt of the evaluation metrics of *Informativeness* and *Overall* in our experiments. The {Language} represents the language of dialogue, while {Input} represents the dialogue that needs to be evaluated.